SEQUENCE ANALYSIS OF FAMILIAL NEURODEVELOPMENTAL DISORDERS

by
Joseph Mark Tilghman

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
December 2020

Abstract:

In the practice of human genetics, there is a gulf between the study of Mendelian and complex inheritance. When diagnosis of families affected by presumed monogenic syndromes is undertaken by genomic sequencing, these families are typically considered to have been solved only when a single gene or variant showing apparently Mendelian inheritance is discovered. However, about half of such families remain unexplained through this approach. On the other hand, common regulatory variants conferring low risk of disease still predominate our understanding of individual disease risk in complex disorders, despite rapidly increasing access to rare variant genotypes through sequencing. This dissertation utilizes primarily exome sequencing across several developmental disorders (having different levels of genetic complexity) to investigate how to best use an individual's combination of rare and common variants to explain genetic risk, phenotypic heterogeneity, and the molecular bases of disorders ranging from those presumed to be monogenic to those known to be highly complex.

The study described in Chapter 2 addresses putatively monogenic syndromes, where we used exome sequencing of four probands having syndromic neurodevelopmental disorders from an Israeli-Arab founder population to diagnose recessive and dominant disorders, highlighting the need to consider diverse modes of inheritance and phenotypic heterogeneity. In the study described in Chapter 3, we address the case of a relatively tractable multifactorial disorder, Hirschsprung disease. We identified new risk genes in a relatively small cohort (190 probands) by combining statistical genetics with functional assays. We then used both known and novel genes

and loci for quantifying individual genetic risk for Hirschsprung disease, on the basis of common and rare variant genotypes. In the fourth and final chapter, we address the case of a highly complex and heterogeneous disorder, autism spectrum disorder. We investigated the basis for exceptionally high genetic risk in 99 families having multiple females affected with autism, showing their risk originates in part from the same genes responsible for *de novo* risk in simplex families. However lack of significant results in gene discovery indicates that functional and external validation is needed for definitive gene finding even in this cohort characterized by high genetic risk.

Thesis readers:

Dr. Aravinda Chakravarti, PhD (advisor)

Dr. Sarah Wheelan, MD, PhD

Preface:

In my graduate work, I have relied primarily datasets and patient cohorts that predated the beginning of my time in graduate school, and the quality and design of these has been fundamental to my success. I am exceedingly grateful not only to my former and current coworkers in the Chakravarti Lab and external collaborators, but to many others who preceded me. I am of course also very grateful to the many affected families, physicians, and genetic counselors who have furnished the collections and clinical phenotyping upon which our work depends.

Foremost, I would like to specifically thank my mentor, Aravinda, who has given me many great opportunities and from whom I have learned a great deal. I especially appreciate the lengths to which Aravinda has gone to provide guidance following my transition to being a remote lab member following his move to New York University. In my time in the Chakravarti lab, I have valued the advice and friendship of many. I worked with Dr. Ashish Kapoor in my rotation project when I first began at Hopkins; he was part of the reason I joined the lab, and he was a valuable source of advice for the several years that we worked together in the lab. I have also especially valued the friendships of Drs. Sumantra Chatterjee and Michael Chou and of Nan Hu.

The Human Genetics program has provided me with an excellent setting in which to complete my thesis, for which I am very grateful to Sandy Muscelli the program administrator, the program head Dr. Dave Valle, the rest of the executive committee, and the many preceptors who helped me feel welcome and enriched my time at

Hopkins. I also count myself fortunate to have benefited from such an excellent peer group in the Human Genetics program and in the School of Medicine as a whole.

I would like to thank Drs. Anthony Leung, Dan Arking, and Rick Huganir for their helpful input and support as members of my thesis committee, and I thank Sarah Wheelan for her support, for serving as chair of my thesis committee, and for serving as a reader for my thesis.

I would also be remiss not to acknowledge the critical role that my undergraduate mentors played in the formation of my identity as a scientist. The combined support and scientific responsibility that I was given as an undergraduate by Drs. David Marsh and Fiona Watson is something that I had never imagined possible as I undertook my undergraduate career. The support I had in the Biology department and Washington and Lee was truly amazing, even outside of my research mentors. And my great experience in Dr. Stephen DiFazio's lab at West Virginia University, doing genetics research for the first time, was what inspired me to embrace genetics as lens through which to better understand biology.

Of course, I am also very grateful for the support and constant arguing of my brothers, parents and the rest of my family, which prepared me well for a career in science. As I have worked to complete my thesis, the support of Dr. Yiwen Dai, my wife, has been critical. And, in the current crisis, I am eternally grateful for my son's several hour naps each afternoon.

Table of Contents

List of Tables

List of Figures

Chapter 2 Figures:

Chapter 3 Figures:

Chapter 4 Figures:

Chapter 1: Introduction

The focus of this dissertation is to understand the utility of exome sequencing for gene and variant identification in three classes of neurodevelopmental disorders – a putatively monogenic syndrome, a relatively tractable multifactorial disorder for which some major genetic risk factors are known, and a highly complex and etiologically heterogeneous multifactorial disorder for which little genetic risk has been explained. To address the case of monogenic syndromes, I present gene discovery using exome sequencing of probands in four syndromic neurodevelopmental disorders within an Israeli-Arab founder population where a single gene and single variant is expected. To address the case of a relatively tractable multifactorial disorder, I present our work on identification of new risk genes and the use of both known and new genes and loci for quantifying individual genetic risk for Hirschsprung disease, where both multiple genes and multiple variants are expected. To address the case of a highly complex and heterogeneous disorder, I investigate genes underlying the exceptionally high genetic risk in families with multiple females affected with autism, a neurodevelopmental disorder with high genetic complexity involving rare and common variation in hundreds of genes. Thus, this dissertation addresses the challenges inherent in exome and or genome sequencing from monogenic to highly complex disorders.

For any genetic disorder, we must ultimately be able to accurately identify genetic risk factors, predict their genotypic risk in the broader population, and understand their role in the disease process. While identification of disease associated genes without a corresponding understanding of their contribution to disease risk in the

general population can aid our understanding of their role in the disease process, accurate estimation of the risk they impart is needed to achieve more complete molecular diagnoses, for risk prediction in the general population, and for understanding the relative contributions of different genes to disease etiology in order to better inform investigations of the disease process.

For monogenic disorders, in which molecular diagnosis is comparatively straightforward, identification of disease-causing variation continues to present many difficulties (Chong et al., 2015). Firstly, there are technical issues which can frustrate the accurate identification of coding variation underlying monogenic disorders. Even when the whole genome is accurately sequenced, our lack of a thorough understanding of genomic function leads us to rely on imperfect genomic annotations that exclude some causal variants from consideration, especially if they cause disease through a hypomorphic, gain of function, or noncoding regulatory mechanism (Eilbeck et al., 2017). In practice, the whole genome sequence of an individual, moreover, is never known perfectly or completely. For example, variants may not be detected or accurately identified using sequencing; somatic mosaic variants that often underlie dominant disorders are especially difficult to accurately identify. Moreover, the identification of monogenic disorders often relies heavily upon careful phenotyping to narrow down potential causative genes, which becomes increasingly difficult when many identical or very similar phenotypes can be caused by different genetic disorders. Narrowing down the cause of a disorder also becomes more difficult when there is only one affected individual in a family, making determination of the inheritance pattern more difficult.

Perhaps the greatest complication in identifying monogenic disorders is that they may not follow a strictly Mendelian inheritance pattern.

While there are some variants in some genes that result in completely penetrant disorders, there are many "monogenic" disease alleles showing incomplete penetrance and variably expressivity. Evidence of this is the presence of many apparently disease-causing genotypes within populations of healthy individuals (Xue et al., 2012). Further, a genetic variant may show Mendelian inheritance on one genetic background but complex patterns in another. Even within families, clinically affected individuals can encompass a wide phenotypic spectrum (Chong et al., 2015), with some mutation bearers not being clinically affected. Ascertainment biases owing to our choosing only the 'Mendelian' cases to examine may contribute to a bias in which we discount exceptions to Mendelian inheritance (Chakravarti, 2011). Therefore, there is the distinct possibility that as we observe more and more individuals, especially those not ascertained on the basis of segregation in families, we may recognize that many Mendelian disorders are more complex than they initially appeared. Thus, addressing the role of sequencing for gene discovery in complex disorders is the main problem.

There are a few key factors that contribute to incomplete penetrance and variable expressivity of disease genotypes. First, even within genes that show Mendelian inheritance, there can be many variants of lesser effect that may result in increased disease risk but are not recognized. Interpreting the pathogenicity of missense alleles, for example, is a common problem which does not always have a yes or no answer, whether one is trying to explain a phenotype or trying to determine genetic risk for the

purpose of counselling. Second, variation in one or many genes may modify the presentation of monogenic diseases, as is the case with cystic fibrosis (Drumm et al., 2005), or may even shift the phenotypic spectrum of a disorder from a disease to a non-disease state. Third, measurable environmental factors are often important modifiers of phenotype. Fourth and lastly, inherent developmental stochasticity can have important roles in determining an individual's ultimate phenotype. Such stochasticity can result from somatic mosaicism of *de novo* variation observable in an individual, or stochastic gene expression and post-transcriptional regulation early in development through gene regulatory networks (Honegger & de Bivort, 2018). Appreciation of the factors that affect penetrance and expressivity of genotypes can have important consequences for genetic risk prediction and for understanding which aspects of the disease process might be targeted to either prevent or treat the disorder.

In the first study I present here, that of putatively consanguineous patients from an Israeli Arab founder population with syndromic disorders, we expect that many of the variant interpretation difficulties I have mentioned will be mollified. In such populations, decreased genetic and environmental variation are expected to make disease genotypes more penetrant because the genetic background is more uniform, and, in such families, high rates of homozygosity are expected to result in a preponderance of homozygous recessive disease, simplifying the variant filtering process by allowing the gene search to be limited to autozygous regions. High diagnostic yields that have been achieved for Mendelian disorders in consanguineous populations result primarily from searches for recessive genes only, which have resulted in a

diagnostic yield of 70% (Beaulieu et al., 2014) compared to around 50% in outbred populations (Chong et al., 2015) where all modes of inheritance were considered. However, for families in endogamous communities without known proximal consanguinity it is important to consider non-recessive modes of inheritance as well. Eaton et al. (2020) found that only 15% of genetic diagnoses for simplex cases originating from an endogamous population had a confirmed recessive basis whereas 46% of diagnoses resulted from *de novo* variation; however, recessive diagnoses were much more common in multiplex families and recessive inheritance accounted for 75% of all diagnoses for families from endogamous populations. These results highlight the importance of considering additional modes of inheritance for presumed consanguineous populations, yet Eaton et al. did not consider dominant etiologies resulting from incomplete penetrance in parents.

Each of the four patients we studied had a different developmental disorder. Three patients had severe developmental delay in combination with other syndromic features, and one patient had an ectodermal dysplasia with mild developmental delay. In contrast to previous studies of such families, we considered all monogenic modes of inheritance for each patient, including incompletely penetrant dominant mutations, in order to determine which etiologies could contribute to the high burden of neurodevelopmental disorders in such populations. For three of the four genes, syndromic features greatly aided gene finding. Our findings, especially of an apparent dominant disorder with reduced penetrance in one proband, highlight the benefit of

considering more than recessive or purely Mendelian inheritance-based explanations even for syndromic developmental disorders in founder populations.

The importance of appreciating greater complexity with respect to Mendelian disorders is further demonstrated by the case of congenital nicotinamide adenine dinucleotide (NAD) deficiency disorder. It is caused by biallelic loss of function mutations in three genes encoding essential enzymes in the kynurenine NAD synthesis pathway, resulting in a combination of vertebral, cardiac, renal, and limb defects (VCRL) or in fetal loss (Shi et al., 2017; Szot et al., 2020). Cuny et al. (2020) showed through studies in mice that a variety of other factors influencing primarily maternal NAD biosynthesis can impact fetal NAD levels so as to result in varying incidence and extent of congenital malformation in a manner dependent on genotype and environment. These risk factors include limited maternal intake of the NAD precursors tryptophan and niacin, hypoxia (which results in decreased oxygenase-dependent synthesis of NAD from tryptophan), and potentially maternal and fetal genotype for 17 genes involved in the NAD biosynthesis pathway that could affect fetal NAD levels. The impact of these new environmental and genetic risk factors suggests that NAD deficiency may be the cause of a number of idiopathic congenital malformations and could allow for nutritional intervention for those mothers who harbor a genotype predisposing to NAD deficiency. While such a clear nutritional intervention will likely be rare for genetic disorders, NAD deficiency demonstrates how we may be missing many variants with reduced penetrance and how those variants can be discovered not through testing association

for every genomic locus for every disorder but through seeking to better understand the biology of seemingly Mendelian disorders.

The second disease studied as part of this dissertation, Hirschsprung disease (HSCR), is also one in which putatively Mendelian forms have led to major advances in understanding of HSCR risk in the general population. HSCR is a neurodevelopmental disorder of the enteric nervous system, in which the more severe forms are characterized by autosomal dominant inheritance and the less severe forms by recessive or multifactorial inheritance (characterized by high recurrence risk in the absence of major risk genes), but the variants associated with both forms have incomplete penetrance (Badner et al., 1990). Prior to the study presented here, 17 genes and several chromosomal disorders had known associations with HSCR; chief among the associated genes are those coding for the receptor tyrosine kinase RET (Edery et al., 1994), reduction of which constitutes a dominant form with reduced penetrance, and the G-protein-coupled receptor EDNRB (Puffenberger et al., 1994). *EDNRB* was initially identified by linkage-*cum*-association, resulting from a high frequency mutation in a founder population, where *EDNRB* had combined with other genetic risk factors to cause what was initially assumed to be a recessive form of HSCR in affected families (Puffenberger et al., 1994). The high frequency of the variants implies that affecteds have both heterozygous and homozygous variants. *EDNRB* is, therefore, a great example of why we should not discount the possibility of non-recessive disease in founder populations. In addition to these single genes and chromosomal disorders, four noncoding variants together confer risk that can vary by 30 fold with increasing risk

allele dosage (Kapoor et al., 2015). In the study I present here we combined genotyping of these common variants, rare coding variants in known and novel genes, and genomic copy number variation in 190 individuals to show that genotype specific odds ratios for HSCR vary by a factor of 67, allowing for HSCR genetic risk stratification on the basis of genes discovered through a combination of linkage, association, and sequence-based analyses.

HSCR represents a case in which risk factors identified through a combination of Mendelian and complex disease genetics collectively translate to a major impact on individual genetic risk. Autism spectrum disorder (ASD), on the other hand, represents a case in which there are many known risk genes but where the contributions of these genes to autism liability in the population are not well understood, despite the high heritability of ASD. This is in part because genetic studies of autism have focused on gene discovery through *de novo* mutations in families with only one affected individual (Iossifov et al., 2014; O'Roak et al., 2012; Sanders et al., 2012, 2015; Satterstrom et al., 2020). Such mutations are non-recurrent with low genetic risk, while the high heritability of autism results from inherited variation.

We, in contrast, chose to exome sequence 99 families with multiple females severely affected with autism, which constitute a group of families with exceptionally high genetic risk on the basis of female sex, severity of the disorder, and familial disease (Turner et al., 2015). This approach to autism genetics assumes that ASD in these families results from a greater burden of rare inherited protein coding variation than in other autism families. This is based in part on the positive relationship between

membership in the same three recurrence risk classes represented by FEMFs (female sex, severe disease, and familial disease) and the proportion of individuals with coding mutations discovered in the major risk gene (*RET*) in HSCR (Emison et al., 2010). A preliminary study in fewer female containing multiplex autism families identified *CTNND2*, and suggested *CYFIP1* as autism risk genes, serving as evidence of this increased burden of coding variants (Turner et al., 2015). We confirmed that FEMFs are enriched for putatively damaging coding variation among 28 high confidence autism genes identified through excess *de novo* variation in simplex families (Sanders et al., 2015), confirming that there are commonalities in etiology between those *de novo* contributions to autism and inherited genetic risk in FEMFs, though it is clear that functional follow up on genetic findings is necessary to confirm our genetic findings in families.

Mendelian disorders with syndromic presentations are overrepresented in human genetics research and especially in clinical genomic sequencing studies because the high penetrance of such disorders increases our power to discover new genes that have relatively simple clinical interpretations. However, in studying such disorders, we must keep in mind that genes discovered in collections of a few exceptional Mendelian families will ultimately need to be interpreted in everyone. As genomic sequencing becomes prevalent, we need to look at the impact of variants in representative, heterogeneous populations to be able to predict genetic risk in individuals accurately, even for variants that have large impacts on risk in discovery populations. Ultimately, if we wish to explain genetic risk in families, it is important to be able to query the impact

of genetic variants on risk across individuals having different genetic backgrounds, different combinations of rare risk alleles, sex, family history, and other underlying disease susceptibilities, which highlights the importance of thoroughly phenotyped population-based cohorts to confirm discoveries made in exceptional populations.

Chapter 2: Identification of Mendelian disease Genes in an Israeli-Arab Community Characterized by Consanguinity

2.1 Chapter Introduction

Clinical exome sequencing in families with Mendelian disorders is successful roughly half of the time (Chong et al., 2015). However diagnostic yield of clinical exome sequencing varies from 8 to 70% based on the inheritance of phenotypes being studied and the sequencing regime applied (Wright et al., 2018). The highest diagnostic yield for exome sequencing, 70%, has been reported for individuals from consanguineous unions or from endogamous founder populations (Beaulieu et al., 2014). The gene finding approach used to achieve this high diagnostic yield was based on finding of recessive genes only. Interestingly, though 11% of the successes in recessive gene finding in the study by Beaulieu et al. resulted from compound heterozygote genotypes rather than from autozygous variants, non-recessive modes of inheritance were not examined. Thus, even many cases of recessive traits even in inbred families do not result from identity by descent. In the case of families from endogamous communities but without known proximal consanguinity, it is especially important to consider non-recessive modes of inheritance. Eaton et al. (2020) found that only 15% of genetic diagnoses for simplex cases originating from an endogamous population had a confirmed recessive basis for disease, with 46% of diagnoses resulting from *de novo* variation. Not unexpectedly, when multiplex and simplex families were taken together, however, recessive inheritance constituted 75% of all diagnoses for families with presumed

consanguinity. These results highlight the importance of considering additional modes of inheritance for presumed consanguineous families, especially for simplex cases. Eaton et al. did not consider dominant etiologies resulting from incomplete penetrance in parents, which should also make substantial contributions to disease in endogamous populations just as they do in all populations. In this study, our goal was to assess the results of exome sequencing in an endogamous founder population by looking for all highly penetrant disease alleles, when no autozygous disease gene variant was identified. We expect that consideration of additional modes of inheritance will ultimately increase diagnostic yield, although it introduces new challenges in interpretation.

In order to test this more complete approach to exome sequencing analysisdisease gene discovery in families from an endogamous founder population, we analyzed exome sequencing from four simplex developmental disorder probands originating from an Israeli-Arab founder population. Each of the four patients we studied had a different developmental disorder with syndromic features, which we utilized, in combination with exome sequencing, to determine their molecular genetic diagnosesbases. Through this study, we hoped to both provide diagnoses for patients, to determine which etiologies contribute to the high burden of neurodevelopmental disorders in founder populations, and to evaluate the potential effectiveness of proband-only exome sequencing for diagnosis in a founder population. Here, I present the general exome sequencing methods I used for diagnosis, followed by each of the

four cases as a separate case studiesy, and I end with a brief discussion of our complete findings across the four cases.

2.2 Exome Sequencing and Analysis Methods

For each of the four patients, exonic sequences were enriched using Nextera Rapid Capture Expanded Exome Kits (Illumina, San Diego, CA). Sequencing was performed on the HiSeq2500 (Illumina, San Diego) instrument with 100 bp paired-end reads. Read alignment to reference genome hg19 (GRCh37) was performed using the Burrows–Wheeler Aligner (Heng Li, 2013) and variant calling was performed using Samtools (Heng Li et al., 2009). Annotation of variants was performed using ANNOVAR (Wang et al., 2010) in combination with in-house scripts.

Following alignment and variant calling, we removed variants with a PHRED quality score less than 30 or coverage less than 10 reads to exclude low quality variant calls. Following variant annotation, we filtered variants to only those with an allele frequency less than 1% in 1000 Genomes (www.1000Genomes.org) and the Exome Sequencing Project (http://evs.gs.washington.edu/EVS/) databases, and which were considered to be damaging based on their predicted genic impact and phylogenetic conservation. For single nucleotide variants (SNVs), we retained splice site, stop-gain, stop-loss, and non-synonymous variants deemed most likely to affect protein function (i.e., phyloP or GERP++ phylogenic conservation score greater than 4 and polyPhen structural prediction of 'damaging'). We also examined small insertion deletion (indel)

variants, which were filtered by SIFT Indel (Hu & Ng, 2013) for those predicted as damaging with confidence score greater than 0.8.

When the disorders were considered likely to be recessive, based on known parental consanguinity, we further filtered to include only homozygous and compound heterozygous variants that were - in a homozygous state in the exomes of other individuals from the same population who had different developmental disorders. However, we went on to investigate other modes of inheritance if a diagnosis was not obtained under the assumption of recessive inheritance. For the one family where there was no known history of consanguinity, we did not include this recessive filtering step, instead relying more heavily on clinical features in order to narrow down candidate genes.

2.3 Case 1: A *PIGN* Mutation Responsible for Multiple Congenital Anomalies–Hypotonia–Seizures Syndrome 1 (MCAHS1) - reprinted with permission from *Am J Med Genet A*

Introduction:

MCAHS1 (OMIM 614080) is an autosomal recessive disorder characterized by developmental delay, hypotonia and epilepsy, combined with multiple congenital anomalies owing to mutations in the *PIGN* gene (Maydan et al., 2011; Ohba et al., 2014). *PIGN* is one of more than 20 genes involved in the glycosylphosphatidylinositol (GPI) anchor biosyntheticsis pathway, of which PIGN protein controls the addition of phosphoethanolamine to the first mannose in GPI (Freeze et al., 2012; Hong et al.,

1999). Mutations in *PIGN*, and seven additional genes involved in GPI biosynthesis, have been identified in individuals presenting with varied neurological abnormalities (Brady et al., 2014; Freeze et al., 2012; Hansen et al., 2013; Krawitz et al., 2013; Kvarnung et al., 2013; Maydan et al., 2011; Ohba et al., 2014). We here report on a girl with MCAHS1 who was born to consanguineous parents and harbors a novel homozygous novel c.755A>T *PIGN* mutation. Our family is the second consanguineous Israeli–Arab family, and the fourth family, reported to date with MCAHS1 resulting from a mutation in the *PIGN* gene (Brady et al., 2014; Maydan et al., 2011; Ohba et al., 2014).

Proband Specific Methods:

Sanger Sequencing: As a confirmation of exome sequencing results, sequence analysis of exon nine of *PIGN*, using genomic DNA from the patient, two healthy sisters, one healthy brother and their parents was performed by amplification of a 229 bp fragment containing the putative mutation deleterious variant identified through exome sequencing. The sense 5'-AAGCATTTCAGAAGTTACTG-3' and the antisense 5'-AAGACATCTAATCCTCTCAA-3' primers were used under the following PCR conditions for DNA amplification: denaturation at 94°C for 5 min; 35 subsequent amplification cycles performed at 94°C for 30 sec, at 55°C for 45 sec and at 72°C for 30 sec; and followed by another 72°C for 5 min. The sequencing reaction was performed using the BigDdye terminator kit and analyzed by the ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems, Warrington, UK), according to the manufacturer's instructions.

Flow Cytometry (FC) Analysis: We examined the effect of the c.755A>T mutation in PIGN on the surface expression of GPI-anchored proteins by staining granulocyte cells with fluorescently-labeled inactive toxin aerolysin (FLAER–ALEXA) (CEDARLANE, Burlington, NC), and mouse antidecay accelerating factors CD16,CD18, CD24, and CD45 antibodies (BD Biosciences, Franklin Lakes, NJ). The antibodies against the following were FITC, PE, APC, and PerCP fluorescently labeled, and used in comprehensive four-color multiparameter Flow Cytometric Analysis on a BD FACSCalibur (BD Biosciences).

Clinical Summary:

Family history: The parents of our proband are first cousins of Israeli–Arab origin. They have two sons and two daughters who are healthy. They had two prior spontaneous abortions and reported another male baby, who died at age of 12 days, following the diagnosis of diaphragmatic hernia (no further details are available). The mother had a brother who died during his 1st year of life but no further details are available. Otherwise, the family history is unremarkable (Figure 1).

Clinical description of the proband: The pregnancy was normal, with no known teratogenic exposure; the mother was 32 years old. The proband (IV4 in Figure 1) was born at term with normal birth weight of 3300 g. Weakness of muscles was noticed at age four months. At age nine months a metabolic work-up that included complete blood count, serum routine chemistry, glucose, lactate, ammonia, biotinidase, creatine kinase, acylcarnitines, amino acids, very long chain fatty acids, and isoelectric focusing of transferrins and urinary organic acids profiles were all normal. Cerebrospinal fluid

analysis for cells, glucose, protein, lactate, and amino acids were normal. Enzymatic assays in white blood cells of enzyme activities for various lysosomal diseases including GM1 and GM2 deficiency, Krabbe, and MLD were negative. Clinical examination showed hypotonia and dysmorphic features “reminiscent of Down syndrome” (Figure 2A,B). At age six months the 1st seizure was noticed; at age 10 months the diagnoses of convulsions, developmental delay, and hypotonia were recorded. EEG analysis revealed epileptiformic bursts. Brain MRI at nine months revealed widening of the sub-arachnoidal space in the frontal and temporal lobes, the lateral ventricles widths being enlarged to 12mm. These findings were interpreted as brain atrophy. Echocardiography and ophthalmological examination were normal. Peripheral blood analysis of karyotype was 46,XX and normal, while FISH analyses specific to chromosomal regions 21q22.3 and 9q34 were also normal. Cytogenetic array CGH yielded no known pathogenic copy number variants (CNV).

Neurological assessment at age 13 months indicated DQ of 40, developmental delay and convulsions. At age 13 months she has had started to roll over, but there was no crawling, sitting or standing. Her weight and height were normal for her age. Her head circumference was 44.5 cm (10–25 percentile). Clinical examination revealed some unusual findings, including, brachycephaly, flat face, up-slanting palpebral fissures, synophrys, squint, large cheeks, small nose and mouth, relatively small ears (4 cm, 2nd percentile), hyperfolded and coarse helices, short neck, and dimples of elbows (Figure 2A). There were no transverse creases of palms and the 5th fingers appeared almost normal. There was general mild hypoplasia of distal parts of all fingers. Brain MRI at age

23 months was interpreted as progressive white matter disease. At age five years there was only partial response to combined anti-convulsive therapy. She was not ambulant, had no speech, and she needed assistance with daily life activities including eating. Ophthalmological examination showed intermediate esotropia, nystagmus with vertical component, blepharitis, and normal fundi. Gastro-esophageal reflux was diagnosed. Clinical examination documented brachycephaly, some hypopigmented macules over the leg, open mouth, and drooling. There was good control of the head but hypotonia of the upper body with postural kyphosis while sitting. There was reduced strength of upper body, but in the lower body there was proximal weakness and increased tone distally. Tendon reflexes were mildly increased, with bilateral clonus. Her growth parameters were normal, with penciled eyebrows and epicanthal folds. The palpebral fissures were up-slanted, the eyes were deep set with nystagmus. We also noted a small nose and somewhat small auricles (5th percentile). Palm and finger length were normal, but the fingers appeared tapering in shape with hypoplastic fingernails; the thumb appeared unusually sharp (Figure 2C). There were prominent blood vessels over the skin, and there was indentation of the middle part of the chest.

Results:

Exome: After filtering exome sequencing data for a recessively inherited disease as described in the exome sequencing methods, 6 SNVs were retained across four genes (*NBPF10, TEP1, CDC27*, and *PIGN*). Only one gene, *PIGN*, is known to be associated with morbidities in OMIM database. The *PIGN* variant is a homozygous mutation located at

Chr18:59814254A>T, c.755A>T, p. D252V. We retained 24 indels in 24 different genes after filtering, but none of these genes had an apparent connection to the phenotype on the basis of reported morbidities in OMIM.

Sanger Sequencing: The mutation was validated by Sanger sequencing which confirmed homozygosity for the c.755A>T variant in the proband IV4. The mother III1, father III2, and one sister IV3 were heterozygous for the same variant. One brother IV1 and one sister IV2 were homozygous for the wild-type allele (Figure 3).

Flow Cytometry: In order to examine the effect of the c.755A>T mutation on the function of *PIGN*, the surface expression of GPI-anchored proteins on granulocytes were analyzed by flow cytometry. Granulocytes were gated after staining with mouse anti-CD45 PerCp to allow us to perform the FC analysis only on blood granulocytes (Figure 4A). The overall expression of GPI-anchored proteins was significantly decreased to 53% of normal levels as revealed by FLAER expression on patient granulocytes (Figure 4B). CD24 and CD18 expression on granulocytes was decreased to 44% and 14% respectively (Figure 4C, D). No abnormal expression of CD16 on granulocytes was observed (Figure 4E).

Discussion:

The involvement of *PIGN* in MCAHS1 has been previously reported in two families. The first family, of Israeli-Arab origin, included seven affected individuals with a missense homozygous c.2126G>A (p.R709Q) mutation (Maydan et al., 2011). The second family, of Japanese origin, included two patients with compound heterozygosity

for c.808T>C (p.S270P) and c.963G>A variants (led to aberrant splicing, in which two mutant transcripts with premature stop codons p.E308Gfs*2 and p. A322Vfs*24 were generated) (Ohba et al., 2014). A third family of North African origin with a splicing homozygous mutation c.1574+1G>A in *PIGN* has been described with an intrauterine phenotype associated with diaphragmatic hernia (Brady et al., 2014).

Here, we describe another family, the second of Israeli–Arab origin, with a *PIGN* mutation. The novel mutation c.755A>T that was detected in our proband was predicted to be "probably damaging" with a score of 1 (polyPhen-2), "deleterious" with a score of 0 (SIFT) and disease causing with a P-value 1 (Mutation Taster). Contrary to the results published by Ohba (2014), the overall expression of GPI-anchored proteins in our patient blood granulocytes was significantly affected by the mutant *PIGN* compared to control samples, as revealed by the significant decrease in FLAER expression. Only CD24, but not CD16 and CD18, expression was drastically decreased on granulocytes from patients as compared to controls. Similar results for CD24, but not CD16, were reported previously (Ohba et al., 2014). These data support the conclusion that the novel mutation detected in our patient causes major damage to the GPI-anchored protein *PIGN*, thus leading to MCAHS1 in our patient.

The girl we described has marked phenotypic overlap with the previously reported affected individuals, including developmental delay, hypotonia, epilepsy, and nystagmus (Maydan et al., 2011; Ohba et al., 2014). However, our proband did not present with congenital anomalies of the cardiac, urinary or gastrointestinal systems (excluding gastro-esophageal reflux) as in other patients. These phenotypic differences

could arise from allele specific effects, involvement of genetic modifiers or be developmental chance effects. The dysmorphic phenotype can be compared with only two of the families previously described, and overlap is present, in particular with respect to the unusual auricles and tapering fingers that were described in at least one individual (Maydan et al., 2011).

In contrast, using exome sequencing, a homozygous splicing mutation c.1574+1G>A in the *PIGN* gene was identified in a fetus of consanguineous parents of North African descent, with multiple congenital anomalies including bilateral congenital diaphragmatic hernia (CDH) (Brady et al., 2014). These authors suggested that the increased severity of the phenotypic features represented by CDH in the tested fetus is due to the homozygous splicing mutation predicting a truncated protein, in comparison to reports of non-synonymous and splicing mutations which likely produce hypomorphic alleles. The family we describe had a male baby, who reportedly died at age of 12 days, following the diagnosis of diaphragmatic hernia. No DNA sample was available and thus mutation analysis could not be performed. Since the mutation detected in our patient is a nonsynonymous mutation, and speculating that the affected baby IV5 (Figure 1) in our family was homozygous for the same mutation, other hypotheses of environmental and genetic modification need to be considered.

To date, mutations in eight genes (*GIPA, GIPL, GIPM, GIPN, GIPO, GIPT, GIPV*, and *PGAP2*) involved in the GPI biosynthesis pathway have been identified in humans. All the affected individuals involved share clinical features including seizures, cardiac defects, skeletal defects, and dysmorphic features (Almeida et al., 2006; Brady et al.,

2014; Hansen et al., 2013; Horn et al., 2011; Johnston et al., 2012; Krawitz et al., 2010, 2013; Maydan et al., 2011; Ng et al., 2012; Ohba et al., 2014). This suggests that some tissues are more sensitive than others to the loss of PIGN activity during embryonic development. This study strengthens the association between *PIGN* mutation and the intellectual disability–hypotonia–seizures syndrome, and expands the mutational spectrum found in this gene.

2.4 Case 2: An *EDAR* Mutation Responsible for Hypohidrotic ectodermal dysplasia

Introduction:

Ectodermal dysplasia (ED) is a heterogeneous group of disorders characterized by lack or dysgenesis of at least two of ectodermal derivatives, including hair, nails, teeth, or sweat glands (Falk Kieri et al., 2014). Hypohidrotic ED (HED) is the most common form of ED and is characterized by a clinical triad of hypotrichosis (sparse hair), abnormal or missing teeth (anodontia or hypodontia) and deficient sweating (hypohidrosis or anhidrosis) (Falk Kieri et al., 2014). HED exhibits a variety of inheritance patterns. X-linked HED (OMIM 3050100), the most common form, is caused by mutations in the ectodysplasin A (EDA) gene, while mutations in the EDA receptor (*EDAR*) and EDAR associated death domain (*EDARADD*) genes result in both autosomal dominant and autosomal recessive forms (Cluzeau et al., 2011; Falk Kieri et al., 2014). Mutations in WNT10A are have also been found shown to be responsible for various autosomal recessive forms of ED (Cluzeau et al., 2011). Taken together, these four genes

account for 90% of the genetic etiologies of hypohidrotic or anhidrotic ED (Cluzeau et al., 2011).

Clinical Summary:

The proband is a female and was referred to the genetics clinic at age 2.2 years due to lack of tooth development, sparse hair, and overheating episodes during physical exercise, suggestive of ectodermal dysplasia (ED). She was born by vaginal delivery at 41 weeks gestation following an uneventful pregnancy, the second child of healthy, distantly related parents of Israeli-Arab ancestry. Her older brother was healthy. The family history was unremarkable.

Physical examination revealed normal growth parameters: her weight and head circumference were 13 kg (between 25th and 50th percentiles) and 47 cm (25th percentile), respectively. She had sparse, light-colored, dry, fine scalp hair. Hypotrichosis was also noted in the eyebrows and eyelashes. Her skin was dry with patchy eczematous regions; neither fingernail nor toenail abnormalities were observed. She had slight coarsening of her facial features, including a wide nasal bridge, low inserted columella, everted vermilion of the lower lip, and large ear lobes. The philtrum was short and smooth, and the upper lip was thin. Additionally, deeply set eyes, narrow and upslanted palpebral fissures, and inflamed conjunctiva were present. She also had nasal speech, anodontia, triangular posterior notched cleft-palate and absent uvula. Borderline delay of motor milestones was present, but speech delay was more marked: at age 2.2 years, she spoke only a few words.

Follow-up examinations at age 4.5 years emphasized her distinctive facial features, which were more marked than in her previous examination, in particular her broad nose, thick lower lip, and hypertelorism. Additionally, premaxillar hypoplasia, short hard palate and a mild degree of glossoptosis were noted. Her nasal speech remained unchanged. Her growth parameters were normal, between the 25th and 50th percentiles (height - 105 cm, weight - 17 kg, head circumference - 48 cm). Prominently, there was a significant improvement of her dermal features, including scalp and eyebrow hair texture and density, and skin dryness. Yet fine scalp hair and mild hyperkeratosis of the palms with subtle hyperpigmentation over fingers joints were observed. Orthodontic assessment showed that she was missing all teeth but the two maxillary central incisors. Developmental milestones were within the normal range for her age. Chromosomal analysis revealed a normal female 46,XX karyotype.

Results:

After filtering variants under the assumption of recessive inheritance as described previously, 5 homozygous SNVs were retained across 5 genes (*CDC27, EDAR, STEAP3, CBLB,* and *SLC22A1*). Only *EDAR* and *STEAP3* have listed morbidities in OMIM, and EDAR dysfunction is known to cause ectodermal dysplasia. There were also 7 homozygous indels across 7 genes (*OR52B4, SMPD1, ATN1, IFI27, C14orf180, TPSD1,* and *MUC20*) of which *SMPD1* and *ATN1* have listed morbidities in OMIM, but none of these had an apparent connection to the clinical phenotype of the proband. The *EDAR* variant is a homozygous C>T mutation located at Chr2: 109546673 (c.77C>T, p. A26V).

The observed allele is unique in that it affects the canonical alanine of the signal peptide cleavage site, which might reduce or prevent cleavage of the peptide and formation of a mature EDAR protein (Z. Zhang & Henzel, 2004). To our knowledge, there have been no previously reported *EDAR* variants affecting this signal peptide.

Discussion:

The *EDAR* gene is 94.9 Kb in size, is located on chromosome 2q11-q13, and contains 12 exons (Cluzeau et al., 2011; Falk Kieri et al., 2014). EDAR protein is a member of the tumor necrosis factor (TNF) receptor family and is activated by its ligand EDA. EDAR uses EDARADD as an adaptor, via its death domain, to build an intracellular NFκB signal transduction complex, which is crucial for normal development of ectodermal organs (Sadier et al., 2014). 22 causative mutations have been reported in the *EDAR* gene, compared to more than 100 mutations in the X-linked *EDA* gene (Azeem et al., 2009; Sadier et al., 2014).

In addition to typical features of ED, our proband exhibited some phenotypic features that have not (to our knowledge) been described before in ED: these are posterior cleft palate and improved hair growth over time. Cleft lip/palate is associated with ectodermal dysplasia in two well characterized genetic disorders: P-63 associated ectodermal dysplasias (EEC Syndrome, AEC Syndrome, Rapp-Hodgkin Syndrome, Limb-mammary Syndrome) and Zlotogora-Ogur syndrome/CLEPD1 caused by mutations in *PVRL1*. Cleft palate has been infrequently described as a part of the clinical picture in ED in several case reports and series (Goyal et al., 2015; More et al., 2013). High-arched

palate, however, occurred in 68% of cases in one study of 19 patients with ectodermal dysplasia (More et al., 2013). Our patient may, therefore, represent an extreme form of the relatively common palatal malformations in ectodermal dysplasia. P63 associated ED, is a group of allelic disorders, caused by heterozygous mutations in the *TP63* gene. It is characterized clinically by a mixture of HED and skeletal malformations, and, in a minority of cases, intellectual disability/mental retardation. P63 expression in the ectodermal surfaces of the limb buds, branchial arches and epidermal appendages in mouse embryos, supports its crucial role in their organogenesis and explains the phenotype of P63 associated ED (Ray et al., 2004; Yang et al., 1999). In addition several cases of non-syndromic cleft lip/palate have been found to be associated with TP63 mutations, supporting the importance of P63 specifically in palatogenesis (Leoyklang et al., 2006; Scapoli et al., 2008).

Zlotogora-Ogur syndrome/CLPED1(Cleft Lip/Palate-Ectodermal Dysplasia Syndrome) is an autosomal recessive condition, characterized clinically characterized by cleft lip/palate, hydrotic ED, developmental defects of the hands, and in some cases intellectual disability (Suzuki et al., 2000). It is caused by mutations in the *PVRL1* gene, which encodes Nectin-1. Nectin-1 is part of the cadherin-based cell-to-cell adherens junctions through its binding to 1-afadin (Suzuki et al., 2000). In a developing mouse embryo model, *PVRL1* mRNA was primarily expressed in the medial edge epithelium of the palatal shelves, the ectodermal component of tooth buds, and the skin surface epithelium (Suzuki et al., 2000), consistent with the phenotypic abnormalities

characteristic of this syndrome. *PVRL1* has more recently been reported as a candidate gene in non-syndromic cleft lip/palate cases (Cheng et al., 2012).

To our knowledge, this is the first report of cleft palate in an *EDAR* associated ED. However, one can imagine how lack of *EDAR* signaling might lead to cleft palate. The EDA-EDAR system, functions by stimulating NFκB -mediated transcription of effectors or inhibitors of the Wnt (Kowalczyk-Quintas & Schneider, 2014; Y. Zhang et al., 2009), sonic hedgehog (SHH) (Kowalczyk-Quintas & Schneider, 2014), connective tissue growth factor (cTGF) (Pummila et al., 2007), fibroblast growth factor (FGF) (Kowalczyk-Quintas & Schneider, 2014), and transforming growth factor beta (TGFβ) (Kowalczyk-Quintas & Schneider, 2014) pathways, regulating interactions within and between epithelial and mesenchymal cells and tissues. These interplays interactions are relevant not only in appendage formation, but also to craniofacial organogenesis and palatogenesis. In a mouse model, failure of posterior muscle development led to posterior cleft palate, explained by secondary loss of TGFβ controlled WNT-β-catenin signaling activity (Iwata et al., 2014). Bone morphogenic protein (BMP) is a protein of the Tumor Growth Factor (TGF) family, with hair and dental placode inhibiting effects (Kowalczyk-Quintas & Schneider, 2014; Pummila et al., 2007). The EDA-EDAR pathway inhibits BMP via expression of various proteins (e.g. connective tissue growth factor, CTGF), enabling placode activation and appendage formation (Pummila et al., 2007). Loss of balance in the BMP pathway, leading to enhanced signaling, has been reported as a cause of complete cleft palate and delayed odontogenic differentiation (L. Li et al., 2013). In addition, absence of BMP inhibitors, such as CTGF and noggin, results in deregulation of

cell proliferation, excessive cell death, and changes in gene expression, leading to complete cleft palate (Goyal et al., 2015).

Our proband exhibited marked improvement in her scalp and eyebrow hair growth, in terms of both texture and higher density, between her first and second evaluations, over the course of 2 years. Although pubertal growth of body hair tends to be normal in ED, spontaneous improvement of scalp and eyebrows hair growth seems rare, and, if found to be frequent in *EDAR* associated ED, might serve as a reassuring prognostic feature when counseling families.

In summary, using exome sequencing, we found a novel homozygous missense mutation in the *EDAR* gene affecting the signal peptide cleavage site, which resulted in autosomal recessive HED with posterior cleft palate and improved scalp hair growth during childhood. We believe that this case provides further insight into the phenotypic spectrum and the natural history of *EDAR* associated ED, which may be under-diagnosed. These findings support meticulous physical examination, and provide some positive prognosis with the possibility of spontaneous improvement over time, at least in some families.

2.5 Case 3: An Interstitial 3p26 Deletion Resulting in Terminal 3p Deletion Syndrome with Incomplete Penetrance

Introduction:

Terminal 3p deletion syndrome (OMIM 613792) has variable phenotypic associations, the most common of these being slow growth, developmental delay,

hypotonia, trigonocephaly, ptosis, hypertelorism, downslanting palpebral fissures, ear and nose abnormalities, and micrognathia. The syndrome is observed in individuals having heterozygous 3pter-p25, 3pter-p26 deletions, or large interstitial deletions affecting this terminal region, with greater severity and multiplicity of clinical features accompanying larger deletions (Shuib et al., 2009). While most individuals with 3p deletion syndrome have *de novo* deletions, some cases are inherited and there have been several cases in which individuals having large 3pter-p25 or 3pter-p26 deletions are apparently healthy and cognitively normal, though they may have severely affected family members with an identical deletion (Pohjola et al., 2010; Shuib et al., 2009; Takagishi et al., 2006; Knight et al., 1995).

As increasing numbers of single gene disorders are discovered near the terminus of 3p, several dominant single gene disorders have been discovered which explain aspects of thise deletion syndrome, several of which can result in intellectual disability. Loss of *SETD5* (Grozeva et al., 2014) and of *BRPF1* (Mattioli et al., 2017) on 3p25 are individually associated with haploinsufficiency-based intellectual disability, yet many features of the syndrome may be observed in individuals with deletions of only 3p26, though with less severe cognitive manifestations.

Clinical Summary:

The proband's parents are first cousins once removed. Retinitis pigmentosa, infertility, newborn death, and motor problems were reported within the extended family. During pregnancy, polyhydramnios, hypospadias, abnormally flexed right foot,

and a small penis were recorded. Penoscrotal hypospadias with chordee was diagnosed at birth. The proband had normal growth parameters at birth. At three months, the proband had moderate hypotonia and DD. At 5 months, the proband had surgery for craniosynostosis.

Exonic sequencing of *FGFR3*, *FGFR2*, and *FGFR1* were normal, as was sequencing of exon 1 of *TWIST1* and *RAB23*. There were abnormal skeletal findings of the pelvis, with an abnormal opening of symphysis pubis.

During a physical exam of the proband at two years, he was thin; his BMI was 2.5 SD below the mean and he had little subdermal fat. His height was in the 10^{th} percentile and his head circumference was in the 25^{th} percentile. The proband had plagiocephaly, curly hair, hypertelorism, squint, high frontal hairline, and small teeth with caries. He had wide, up-turned nares, retro-micrognathia, thin nasal bridge, and a short flat philtrum. He also had an abnormal left nipple.

Results:

After filtering exome sequencing for a recessively inherited disease as described in the exome sequencing methods, 8 variants were retained across 6 genes (*NBPF10, MKI67, KLHL33, ZNF717, CDC27, AMER1*). Only *AMER1*, which has listed morbidities in OMIM, and was a candidate but disruption of *AMER1* would not result in the features observed in the proband. After filtering for indels, 19 variants across 19 genes were retained. Of these genes, 2 have listed morbidities in OMIM unrelated to those observed in the proband.

Array CGH was also performed on this proband, as part of his clinical genetic workup, in order to identify potentially pathogenic CNVs. A 3p26 deletion spanning chr3: 61,891 - 4,602,285 was discovered in the patient and array CGH on the patients parents showed that the deletion was inherited from the proband's father. The deletion was further validated and resolved using the Plink PLINK homozygous run caller in combination with exome sequencing read depth. The CNV identified by array was thus extended slightly by way of exome sequencing to a 4.6 Mb deletion at chr3: 61,891 – 4,683,606. Other individuals with 3p26 deletions have been reported as having a 3p deletion syndrome with comparable clinical features to those observed in the proband. 3p26 deletions are also, however, sometimes observed in apparently healthy individuals. The deletion harbored by the proband results in loss of 10 protein-coding genes, two of which, *LRRN1* and *ITPR1*, are intolerant of to single copy loss of function based on depletion of such variants in gnomAD (Karczewski et al., 2020).

Discussion:

The incompletely penetrant 3pter-p25 and p26 deletions that have been observed show that, while there are many apparently haploinsufficient genes on the terminus of 3p, none of them are completely penetrant. In the deletion present in the proband and in his apparently unaffected father, two apparently dominant disease genes are disrupted.

ITPR1 encodes the inositol 1,4,5-triphosphate (IP3) receptor, which modulates intracellular calcium signaling, and is partially deleted in both the proband and the

father. The phenotype associated with loss of one copy of *ITPR1* is fairly well defined and may have implications for the proband and other family members who may harbor this deletion. Heterozygous *ITPR1* loss of function mutations may result in either congenital non-progressive (OMIM 117360) or progressive spinocerebellar ataxia (OMIM 609958). Non-progressive spinocerebellar ataxia is characterized by delayed motor development, hypotonia, cognitive delay and progressive ataxia, which may explain the hypotonia and developmental delay observed in the proband. It is possible that the same variant, on the father's genetic background, might result in progressive spinocerebellar ataxia in the father, which can remain asymptomatic into adulthood. Thus loss of one copy of *ITPR1* may well provide an explanation for motor problems reported within his extended family and may eventually manifest themselves in the father.

LRRN1 encodes leucine-rich repeat neuronal protein 1, which is involved in the formation of the midbrain-hindbrain boundary in chick development (Tossell et al., 2011). Loss of Lrrn1 in mice, moreover results in behavioral phenotypes (http://www.informatics.jax.org/marker/MGI:106038). Loss of one copy of *LRRN1* does not, however, have a well-defined human phenotype associated with it, as does *ITPR1,* though its important role in brain development and intolerance of variation indicates that loss of a single copy of *LRRN1* may contribute to the features of terminal 3p deletion syndrome.

Neither of the these two haploinsufficient genes provides a complete explanation for the features observed in the patient, which are likely influenced by loss

of the other eight genes deleted in the proband. However the apparent relevance of these two genes to the phenotype observed in the proband highlights the importance of evaluating large genomic deletions based on depletion of loss of function variation in healthy individuals, which can identify disease genes while allowing for incomplete penetrance. This approach is an improvement over past work to identify genes underlying terminal 3p deletion syndrome phenotypes based on the identification of completely penetrant critical regions which do not exist in an incompletely penetrant syndrome.

2.6 Case 4: A missense mutation in the C-terminal zinc finger domain of *ZEB2* Results in Relatively Mild Mowat-Wilson syndrome

Introduction:

Mowat-Wilson syndrome (OMIM 235730) is a moderate to severe ID syndrome characterized by a distinctive facial gestalt, often associated with microcephaly, epilepsy, corpus callosum agenesis, heart defects, urogenital malformations, and Hirschsprung's disease (HSCR). Almost all features outside of the facial gestalt are incompletely penetrant, which is a complicating factor in accurate diagnosis of the disorder. HSCR was initially the basis of case ascertainment, but is now known to be only 60% penetrant (Mowat, 2003). Mowat Wilson syndrome is caused by *de novo* mutations affecting one copy of the *ZEB2* gene, which encodes a transcription factor having important roles in neural crest cell migration and corticogenesis (Seuntjens et al., 2009; Van de Putte et al., 2003). Almost all of the more than 180 reported pathogenic alleles

in Clinvar result in stop-gained mutations within or deletions of all or much of the gene, for which no functional protein is expected (Landrum et al., 2018), with only a handful or reported pathogenic missense changes.

A recent study of the phenotypes and genotypes of 87 patients with Mowat-Wilson syndrome, which included no cases harboring missense alleles in *ZEB2*, found that a less severe presentation was associated with cases where some functional protein was expected (Ivanovski et al., 2018). The study also found that the most consistent clinical associations with Mowat-Wilson syndrome are microcephaly and seizures. Seizures are a risk factor for regression of cognitive and motor skills in Mowat-Wilson syndrome (Bonanni et al., 2017).

Clinical Summary:

The male proband was born to unrelated parents. The mother's nephew (a son of her sister, whose parents are not related) was reported to have DD and hypotonia. The father has a nephew with DD, whose parents are not related.

Brain cysts were observed in the proband during pregnancy. His birth weight was 4.5 Kg (the mother is not diabetic). There was developmental delay, with appearance of speech that regressed. Febrile convulsions appeared at 2 years.

In a physical examination at 3 years of age the proband was noted to share some dysmorphology with his father. At this time, the proband had no speech, a wide gait, stereotypic movements of his hands, and head banging. The proband had normal growth parameters with medial flare of the eyebrows, deep set eyes, epicanthal folds,

flattening of the nose, short philtrum, narrow palate, folded auricles, and prominent earlobes. The proband also had 5 hyperpigmented spots and tapering fingers.

The proband had a normal metabolic work-up, normal brain magnetic resonance imaging and spectroscopy, normal electroencephalogram and normal brainstem evoked response audiometry. Cytogenetic array CGH yielded no known pathogenic copy number variant (CNV). *TCF4* sequencing to screen for Pitt-Hopkins syndrome resulted in no pathogenic findings.

Results:

After filtering exome sequencing as described in the exome sequencing methods, without assuming recessive inheritance, 129 potentially pathogenic variants were retained across 81 genes. Of these 81 genes, 17 have reported morbidities in OMIM. Of those 17 genes, each of which harbored a single heterozygous variant, three genes (*ZEB2, IFIH1,* and *KCNT1*) have reported phenotypes in OMIM that include developmental delay, though the presence of the *IFIH1* and *KCNT1* variants at low rates in healthy controls excluded their being sufficient for disease and, therefore, high-value candidates, in addition to a mismatch of the patients clinical features with those syndromes. The *ZEB2* variant, on the other hand, is novel and Mowat-Wilson Syndrome is sufficient to explain all of the dysmorphic and pathologic features observed in the proband. The *ZEB2* variant discovered is at Chr:145147446G>A, NM_014795 c.C3217T, p.H1073Y in *ZEB2* and it was confirmed as *de novo* by Sanger sequencing in of the parents. After filtering, 64 potentially pathogenic indel variants were across 57 genes

were also retained. Of these genes, 6 have listed morbidities in OMIM, but none are consistent with the features of the proband.

Amino acid position 1073 of ZEB2 encodes a histidine involved in zinc ion chelation within a highly conserved C-terminal zinc finger domain (C-ZF) (http://www.uniprot.org/uniprot/O60315). And it is only one residue away from another C-ZF variant previously reported to cause a relatively mild form of Mowatt-Wilson syndrome (Ghoumid et al., 2013). There were three zinc-finger disrupting mutations reported in that study, all of which prevented binding of ZEB2 to the E-cadherin promoter *in vitro* and rescued cranial morphology and neural crest cell proliferation in zebrafish morphants of a *ZEB2* orthologue to varying degrees.

Discussion:

The three previously reported cases of relatively mild Mowat-Wilson syndrome, caused by missense mutations affecting ZEB2 C-ZF domains were caused by the mutations p.Tyr1055Cys, p.Ser1071Pro and p.His1045Arg (Ghoumid et al., 2013). Like the proband reported here, none of these cases had microcephaly and two of three had no observed brain anomalies. Only one of the three cases had observed brain anomalies, the patient having the p.His1045Arg mutation, and this mutation showed least rescue of zebrafish morphants, perhaps explaining the more severe phenotype of the patient, who also had hypospadias. The primary departure of the case reported here, from these other three other patients having ZEB2 C-ZF domain mutations, is his absence of speech (present to some extent in the three previously reported cases) ,

which may be explained by seizure-associated regression, as the patient is reported to have had regression of early language development in addition to febrile convulsions. Only one of the three previously reported cases suffered from seizures and no regression was reported in any of the three cases.

This work provides additional support for mild presentations of Mowat-Wilson syndrome associated with missense mutations affecting the ZEB2 C-ZF domains, in which few if any of the congenital malformations common to Mowat-Wilson syndrome are present, with the exception of facial dysmorphology. While the three previously described cases show superior language development compared to typical Mowat-Wilson presentations, the case described here stands as an exception, suggesting that the phenotypic heterogeneity observed for individuals having a complete loss of one copy of *ZEB2* is also the case for C-ZF domain mutations.

2.7 Chapter Conclusions

Two of the four probands investigated as part of this project have monogenic recessive disease resulting from autozygosity. Both sets of parents for these two probands have known relatedness to each otherare known relatives – the parents of the proband affected by MCAHS1 are first cousins and the parents of the proband affected by ectodermal dysplasia reported being distantly related. However, in another family where the proband's parents are first cousins once removed, the proband's condition is not explained by autozygous inheritance. A methodology that limited gene finding to only recessive modes of inheritance would therefore have missed the dominantly

inherited 3p terminal deletion syndrome in the proband, as would a methodology based on only recessive and *de novo* filtering, the approach used by Eaton et al. (2020). For the one patient we examined who did not have reported parental relatedness, however, the patient's disorder is explained by *de novo* inheritance of Mowat-Wilson syndrome, consistent with the finding by Eaton et al. that nearly half of simplex cases originating from endogamous populations have *de novo* disorders. We had limited ability to identify *de novo* dominant mutations, as we performed proband-only exome sequencing, but the distinctive facial dysmorphology of Mowat-Wilson syndrome allowed us to narrow down the many potential disease-causing mutations more effectively than might be expected, which emphasizes the importance of combining exome sequencing with a thorough clinical description in order to identify the cause of an unknown developmental disorder. Additionally, trio-based exome sequencing should be routine.

Our findings support the point made by Eaton et al. that, while endogamous founder populations are enriched for recessive disorders owing to increased autozygosity, looking for only recessive disease in these populations will result in missing the dominant disorders that also occur in all populations. Our finding of an incompletely penetrant large deletion also shows the importance of considering incompletely penetrant inheritance in combination with *de novo* inheritance as potential causes of disease in founder populations. Considering incompletely penetrant variants in founder populations is of special importance because we expect some rare damaging variants to exist at relatively high frequencies as a result of founder effects in these populations. Our case of terminal 3p deletion syndrome also highlights the

importance of considering copy number variation in the diagnosis of developmental disorders in founder populations.

While endogamous populations are viewed as important for exome sequencing studies because of the expected high diagnostic yield resulting from increased autozygosity, these populations are subject, like other populations, to both *de novo* and inherited dominant disorders and gene finding must consider these hypotheses when a recessive disease sufficient to cause explain the disorder is not identified. A corollary of the need to screen for *de novo* inheritance is that, when possible, exome or genome sequencing of parents in addition to probands will substantially increase diagnostic yield in founder populations.

2.8 Chapter 2 Figures:

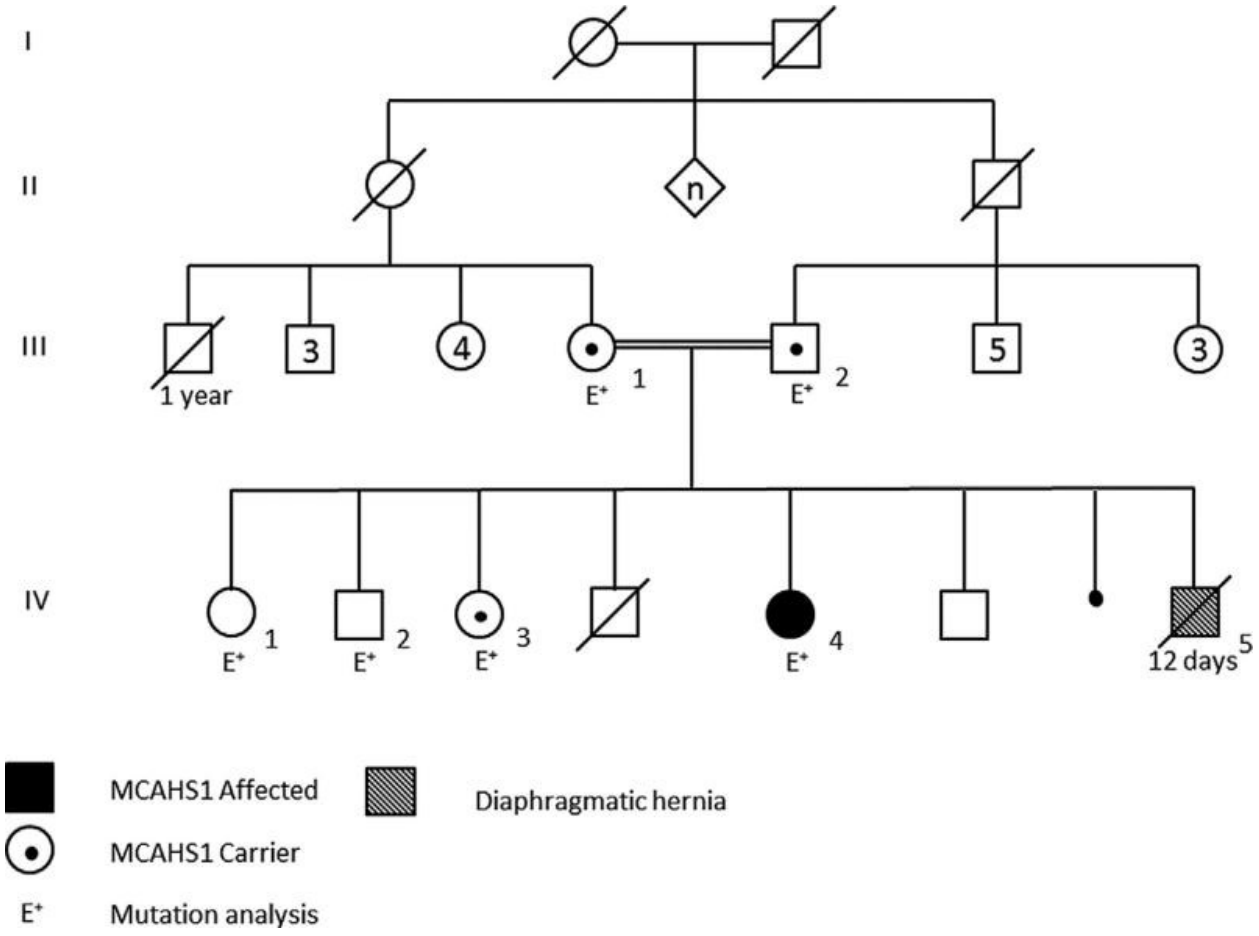


Figure 5: *Pedigree of the Israeli-Arab family presenting with MCAHS1*.

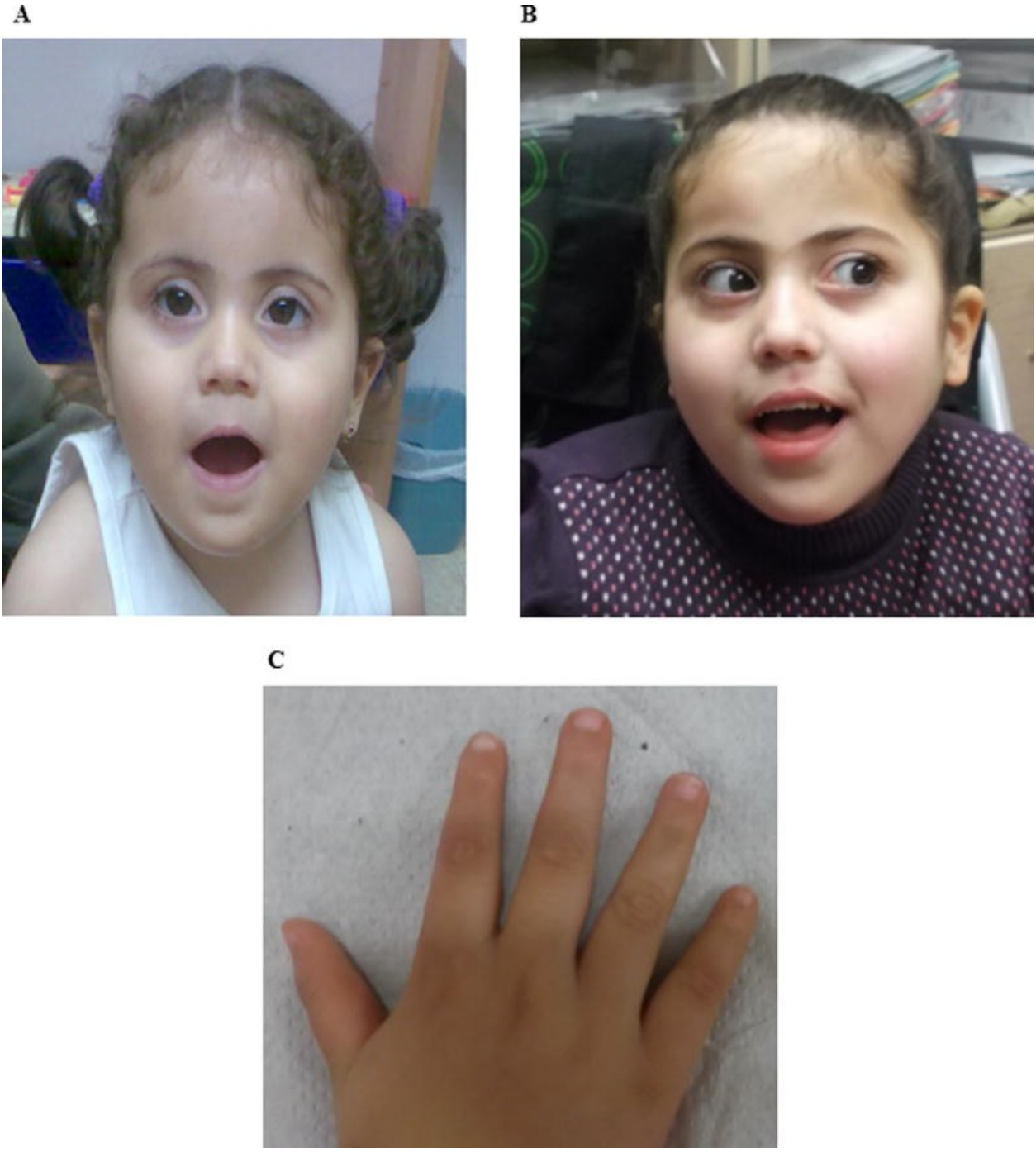


Figure 6: *Photographs of: A. the patient face at age of 1 year and 9 months. B. the patient face at age of 6 years and 4 months. C. the patient hand at age of 5 years and 2 months.*

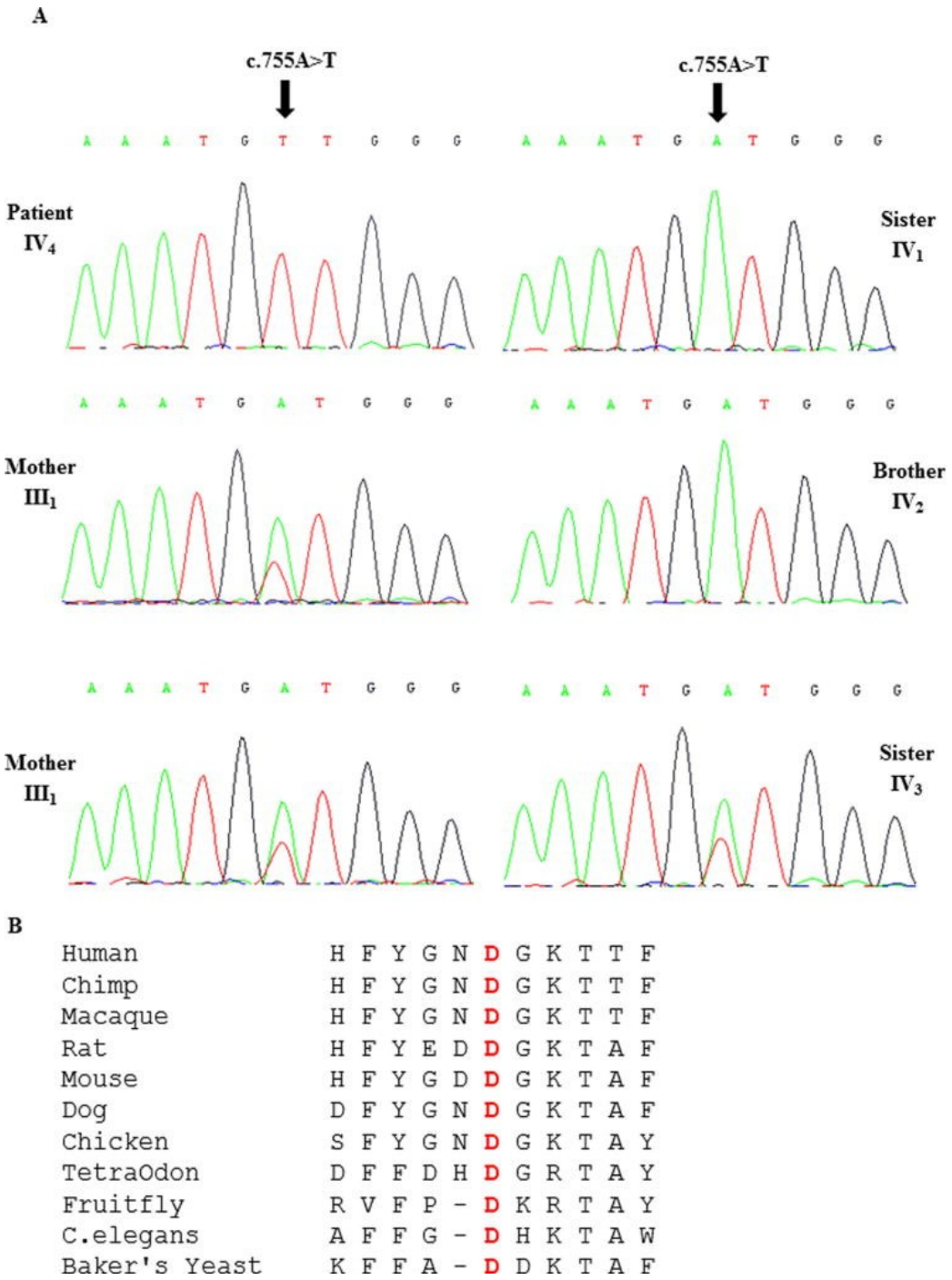


Figure 7: *A. DNA sequence electropherograms of the c.755A>T mutation identified in exon 9 of PIGN in the patient, her brother, her two sisters, and her parents. B. Alignment of different PIGN amino acid sequences with human PIGN. The conserved aspartic acid at position 252 is highlighted in red.*

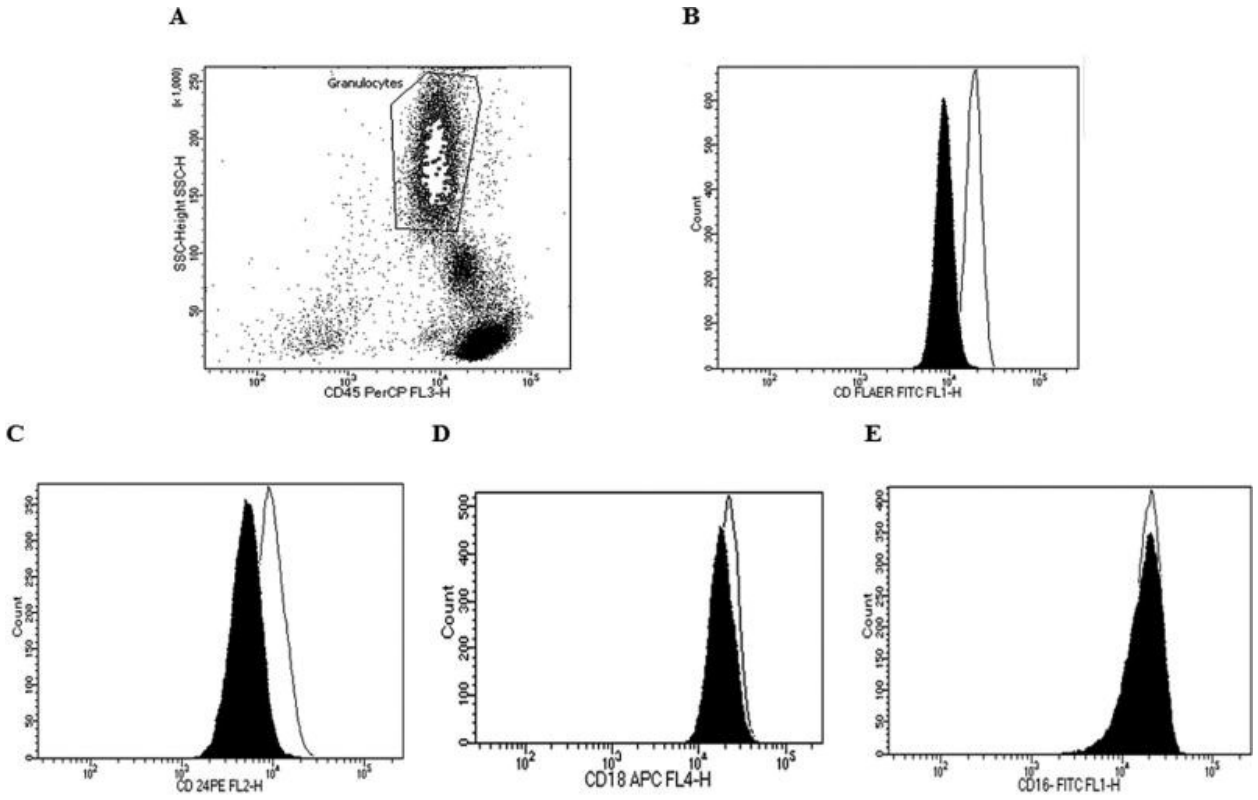


Figure 8: *A. Gated granulocyte cells stained with mouse anti CD45. B. Surface expression of overall GPI-anchored proteins as revealed by FLAER expression on blood granulocytes. C–E. Expression of CD24, CD18 and CD16, respectively, on blood granulocytes. The dark shadow represents our patient, solid line represents normal controls.*

Chapter 3: The Molecular Genetic Anatomy and Risk Profile of Hirschsprung's Disease

3.1 Introduction

Hirschsprung's disease is characterized by the lack of ganglia in the myenteric and submucosal plexuses of the gut. It is a "model" complex disorder because it exemplifies multifactorial inheritance and yet has been molecularly tractable (Amiel et al., 2008; Carter, 1969; Chakravarti & Lyonnet, 2001). The disease (with an incidence of 15 cases per 100,000 live births) is characterized by high heritability (>80%) and marked sex differences (male:female ratio, 4:1) (Chakravarti & Lyonnet, 2001). Patients have aganglionosis affecting bowel segments of variable length, as a result of incomplete rostral-to-caudal enteric neuronal colonization; on the basis of these segment lengths, the condition is classified as short, long, or total colonic aganglionosis (see the Supplementary Appendix, available with the full text of this article at NEJM.org). Approximately 18% of patients have multiple anomalies, some with specific syndromes; approximately 12% have major chromosomal variants (Amiel et al., 2008; Chakravarti & Lyonnet, 2001). Features of Hirschsprung's disease include its high (3 to 17%) sibling recurrence risk (i.e., the risk of being born with the disease, given that one full sibling is affected) and the variation in risk according to sex, segment length, and familiality (Chakravarti & Lyonnet, 2001).

Hirschsprung's disease has multiple causes, although no environmental causes are known (Bodian & Carter, 1963; Passarge, 1967). Complex segregation analyses have

refined this view by showing genetic heterogeneity according to the extent of aganglionosis. The long form is characterized by autosomal dominant inheritance and the short form by recessive or multifactorial inheritance, and the variants associated with both forms have incomplete penetrance (Badner et al., 1990). This finding led to the discovery of 17 genes with approximately 500 rare disease-associated coding variants, chiefly the genes encoding the receptor tyrosine kinase RET and the G-protein–coupled receptor EDNRB (Table S1 in the Supplementary Appendix) (Amiel et al., 2008; Edery et al., 1994; Emison et al., 2005, 2010; Garcia-Barcelo et al., 2009; Jiang et al., 2015; Kapoor et al., 2015; Puffenberger et al., 1994). Four noncoding variants, individually conferring moderate risks (odds ratio, 1.6 to 3.9) but together conferring risk that can vary by as much as a factor of 30 with increasing risk allele dosage (Kapoor et al., 2015), are genetic modifiers of Hirschsprung's disease (Arnold et al., 2009; Emison et al., 2010). These data suggest widespread and variable genetic susceptibility to the disease from multiple genes, reflected in the differing presentations and recurrence risks among relatives.

We suspected that, in contrast to the genetic risk factors for other complex diseases, many genetic risk factors make individually large contributions to the risk of Hirschsprung's disease. We undertook genotyping, exome-sequencing, and functional assays to study pathogenic alleles in a set of patients with Hirschsprung's disease with representative phenotypes. Beyond studying known genes and identifying new ones, we investigated the variation in risk according to the type of pathogenic allele, the contribution of each type of allele to Hirschsprung's disease in the general population,

and the distribution of these types of alleles across phenotypes. Our primary goal was to enable genetic stratification of patients in order to determine how genetic susceptibility manifests in clinical disease and its penetrance. Such genetic stratification could be used to determine whether postsurgical outcome — for example, continued bowel dysfunction and enterocolitis, which is reported in 30 to 50% of patients (Dasgupta & Langer, 2008; Menezes et al., 2006) — is related to genotype.

3.2 Methods

Participants and Genome-wide Analyses:

We conducted exome sequencing of samples from 190 patients of European ancestry and 47 of their affected relatives (7 parents, 12 children, 17 siblings, and 11 second-degree relatives) with diverse phenotypes. The control sample used in exome sequencing consisted of publicly available, ancestry-matched exome data on 740 samples from the 1000 Genomes Project and the National Institute of Mental Health Repository. For the analysis of common noncoding variant studies, we used a different set of 627 control samples that were genotyped in our laboratory: 404 from the 1000 Genomes Project and an additional 223 “pseudo-controls” (generated from the chromosomes not transmitted to the affected child in 254 parent–child trios) (Kapoor et al., 2015). For the analysis of copy-number variants, we used a third control set of 19,584 adults of European ancestry (DePristo et al., 2011).

Pathogenic Alleles, Genes and Loci:

For assessing the effect of common noncoding variants, we used four disease associated SNPs — rs2435357, rs7069590, and rs2506030 in *RET* and rs11766001 in the SEMA3 gene cluster (Chatterjee et al., 2016; Kapoor et al., 2015). We have previously shown that the *RET* noncoding variants are located within transcription enhancers bound by the transcription factors RARB, GATA2, and SOX10; these variants lead to reduced *RET* expression and an elevated risk of Hirschsprung's disease (Chatterjee et al., 2016). Although the causality of the rs11766001 polymorphism in the SEMA3 locus is unproven, considerable data support causality or a strong association with a causal variant in *SEMA3C* or *SEMA3D*, which have been shown to be necessary for gut innervation (Jiang et al., 2015; Kapoor et al., 2015). Coding pathogenic alleles at each gene were defined as nonsense or missense changes in codons encoding amino acids that are conserved (with respect to their position in the oligopeptide) across species, splice-site single nucleotide variants, and all coding insertion–deletion variants with a frequency of 5% or less. These definitions gave acceptable levels of true and false positives at known Hirschsprung's disease genes (Table S1 in the Supplementary Appendix) (Stenson et al., 2014). Disease-associated coding variants can have incomplete penetrance and be present in controls; therefore, we identified Hirschsprung's disease-associated genes as those that had a greater number of unique pathogenic alleles in patients than in controls (Fig. S4 in the Supplementary Appendix). We assessed large copy-number variants (deletions of more than 500 kb and duplications of more than 1 Mb) with a frequency of less than 1% among controls to determine whether they were significantly enriched among patients or had previously

been found to be associated with a developmental disorder (Tables S8 and S9 and Fig. S5 in the Supplementary Appendix) (DePristo et al., 2011; Itsara et al., 2009). Additional details are provided in the Methods section in the Supplementary Appendix.

To assess the role of a gene in Hirschsprung's disease, we first used reverse transcriptase polymerase chain reaction (RT-PCR) to assess its RNA expression in the human embryonic gut at Carnegie stage 22, by which time gut neurogenesis is complete (Fig. S6 in the Supplementary Appendix). Second, we tested gene expression by RNA sequencing and RT-PCR in the developing mouse gut during the equivalent developmental period (embryonic day 10.5) (Fig. S6 in the Supplementary Appendix). Third, we used morpholinos (antisense oligonucleotides) to knock down gene expression in zebrafish embryos at 6 days after fertilization and enumerated the enteric neurons colonizing the gut relative to controls (Fig. S7 in the Supplementary Appendix) (Jiang et al., 2015).

Statistical Analysis:

Population-level risks were estimated for groups of pathogenic alleles, genes, or loci with the use of odds ratios with significance thresholds (corrected for multiple testing) and 95% confidence intervals (Kapoor et al., 2015; Purcell et al., 2007). The odds ratios were converted to estimated population penetrance (equivalent to the population incidence or risk) with Bayes' theorem, under the assumption of an incidence of 15 cases per 100,000 European-ancestry live births (Emison et al., 2010). Allele frequencies among controls were obtained from a variety of public resources (Coe

et al., 2014; DePristo et al., 2011; The 1000 Genomes Project Consortium, 2015) to estimate the population attributable risk. Additional details are provided in the Methods section in the Supplementary Appendix.

3.3 Results

Common Regulatory Variants and Risk:

Four common transcription-enhancer variants were associated with a moderate risk of Hirschsprung's disease in our sample of 190 patients and 627 controls (Table S2 in the Supplementary Appendix) (Emison et al., 2005, 2010; Garcia-Barcelo et al., 2009; Jiang et al., 2015; Kapoor et al., 2015). The frequency of these variants allowed us to estimate their total effect according to dosage in reference to persons with one allele (none had zero alleles): a risk of Hirschsprung's disease (odds ratio >1) is evident only with three or more alleles (Table 1), but, in view of multiple comparisons, the risk was considered significant only when at least five risk alleles were present (odds ratio, 4.54; 95% confidence interval [CI], 3.19 to 6.46; $P = 1.22\times10^{-16}$) (Table 2). Thus, the population risk of Hirschsprung's disease varies by a factor of 24, from approximately 1 case per 19,100 live births (0 or 1 risk allele) to 1 case per 710 live births (seven or eight risk alleles) according to enhancer risk-allele dosage, which shows the wide differences in basal susceptibility to Hirschsprung's disease.

Risk Associated with Rare Coding Variants:

We first tested whether coding pathogenic alleles, as we defined them, for the 17 known Hirschsprung's disease genes statistically discriminated patients from controls

(Table S1 in the Supplementary Appendix). As compared with the 29 pathogenic alleles found in 71 (9.6%) of 740 controls, 36 pathogenic alleles were found in 41 (21.6%) of 190 patients, a percentage 2.25 times as high ($P = 5.97\times10^{-6}$) (Table S6 in the Supplementary Appendix), which indicates a higher burden of pathogenic alleles in patients. Furthermore, the pathogenic alleles that were found in patients had a significantly lower mean frequency in an external reference population, the Exome Aggregation Consortium database (ExAC) (Exome Aggregation Consortium et al., 2016), than did the pathogenic alleles found in controls (5.58×10^{-4} vs. 1.11×10^{-3}, $P = 2.14\times10^{-5}$) (Table S6 in the Supplementary Appendix), which indicates that the rare coding changes observed in patients have been subject to greater purifying selection than those observed in controls. That is, even though pathogenic alleles in both patients and controls met our definition of pathogenicity, when we compared the frequency of each set (variants in the patients being one set and variants in the controls the other) with the frequency of the specific variants of each set in persons in the ExAC database, those of the patient set were less frequent in the ExAC database than were those in the control set.

To assess the enrichment of pathogenic alleles for each gene, we estimated the probability (P value) of finding as many or a greater number of distinct pathogenic alleles in patients, restricting our analysis to 15,963 single-nucleotide variants in 4027 genes for which there was at least one identified pathogenic allele in both patients and controls. We identified 3 genes, *EDNRB*, *ADAMTS17*, and *ACSS2* that exceeded the significance threshold of 1.24×10^{-5} (5% significance across 4027 genes) (Fig. S4 in the

Supplementary Appendix). More broadly, at a P value threshold of 0.001, we found 10 genes instead of the expected 4 ($P = 1.3\times10^{-3}$) (Table 3). We performed functional tests on these 10 genes to distinguish false from true candidates.

The top 10 genes had a minimum of 4 pathogenic alleles each and included both of the major genes, *RET* and *EDNRB*. We also found evidence of 7 novel Hirschsprung's disease genes — *ACSS2*, *ADAMTS17*, *ENO3*, *FAM213A*, *SH3PXD2A*, *SLC27A4*, and *UBR4* — on the basis of both an excess of pathogenic alleles and enteric nervous system gene expression in humans and mice during enterogenesis; assays in zebrafish further confirmed *ACSS2*, *ENO3*, *SH3PXD2A*, and *UBR4* (Fig. S6 in the Supplementary Appendix). The 7 novel genes harbored 39 distinct pathogenic alleles occurring in 40 patients (21.1%), as compared with 23 distinct pathogenic alleles occurring in 28 controls (3.8%) ($P = 3.46\times10^{-16}$). Of the 39 pathogenic alleles in patients, only 6 were identified in 8 controls (1.1%). When all 24 Hirschsprung's disease genes were considered, we identified 75 unique pathogenic alleles occurring in 34.7% of patients (66 of 190), a percentage significantly higher than the 5.0% observed among controls (37 of 740; odds ratio, 10.02; 95% CI, 6.45 to 15.58; $P = 3.41\times10^{-25}$) (Table 2). The mean allele frequencies of the pathogenic alleles in patients and controls in the ExAC database are 4.22×10^{-4} and 8.26×10^{-4}, respectively, a difference similar in magnitude to the difference we observed for alleles in the 17 previously known Hirschsprung's disease genes. The causality of these variants is further confirmed by higher-than-expected genotype concordance between probands with coding pathogenic alleles and their affected

relatives (P = 0.005) (Tables S6and S7 and the Methods section in the Supplementary Appendix).

Pathways and Functional Groups:

Owing to genetic heterogeneity and chance fluctuations, the overall contribution of pathways to Hirschsprung's disease can be estimated more accurately than that of individual genes (Table S7 in the Supplementary Appendix). In Hirschsprung's disease, the RET and EDNRB signaling pathways play major roles with strong epistatic interactions (Edery et al., 1994; Emison et al., 2010; Puffenberger et al., 1994). Thus, we considered members of the RET (*GDNF*, *NRTN*, *GFRA1*, and *RET*) and EDNRB (*ECE1*, *EDN3*, and *EDNRB*) signaling modules for burden analysis. A third pathway, also epistatic to RET, involves the class 3 semaphorins and their receptors: here we consider only *SEMA3C* and *SEMA3D* because of their association with Hirschsprung's disease (Jiang et al., 2015; Kapoor et al., 2015). A fourth class consists of the transcription-factor genes (*SOX10*, *ZEB2*, *PHOX2B*, and *TCF4*) that are critical to the early development of the enteric nervous system and harbor rare coding variants that cause Hirschsprung's disease–associated syndromes (Table S1 in the Supplementary Appendix). We considered two additional categories: other known genes (*KIF1BP*, *L1CAM*, *IKBKAP*, and *NRG1*) (Amiel et al., 2008; Garcia-Barcelo et al., 2009) and the seven novel genes identified in this study.

We compared the total numbers of pathogenic-allele genotypes in each of these six classes or pathways among the 66 variant-positive patients with their corresponding

frequencies among controls (Table S7 in the Supplementary Appendix). Genes encoding members of the EDNRB pathway (odds ratio, 69.03; 95% CI, 8.68 to 547.92), transcription-factor genes (odds ratio, 4.15 to 307.72), and novel genes (odds ratio, 23.2; 95% CI, 11.04 to 48.72) had the largest risk effects, followed by genes encoding members of the RET pathway (odds ratio, 16.03; 95% CI, 5.21 to 49.28) and *SEMA3C* and *SEMA3D* (odds ratio, 2.65; 95% CI, 1.25 to 5.60). Other known genes (odds ratio, 3.15; 95% CI, 1.22 to 8.09) also made measurable risk contributions, but with an order of magnitude smaller effect. These risk rankings were reflected in the inverse contributions of these classes to the total risk of Hirschsprung's disease. Pathogenic alleles causing greater risk probably have higher penetrance and are therefore selected against with greater intensity. If so, the abundant coding variants in genes of the RET pathway have lower penetrance than coding variants in the genes of the EDNRB pathway, the genes encoding transcription factors, and the novel genes.

These data also indicate that *RET* has a smaller coding-variant risk burden than previously believed: 6.3% of the patients (12 patients) had *RET* coding pathogenic alleles, in contrast to approximately 50% from the older data (Chakravarti & Lyonnet, 2001; Edery et al., 1994). This difference could arise from differing definitions of pathogenicity or from the preponderance of familial and severe cases in earlier studies. Nevertheless, *RET* regulatory pathogenic alleles, which have even lower penetrance than coding pathogenic alleles (Emison et al., 2005), were prevalent and, together with *RET* coding variants, conferred substantial risk in 92 of 190 patients (48.4%); this finding highlights the fact that reduced *RET* expression is the predominant cause of

Hirschsprung's disease. Moreover, coding or noncoding (in the case of *RET* transcription enhancer variants) pathogenic alleles in affecting genes that encode members of the RET regulatory network (Chatterjee et al., 2016), which is made up of RET, its transcription factors (RARB, GATA2, and SOX10), its ligands (GDNF and NRTN), and its coreceptor (GFRA1), were found in 120 of our patients (63.2%). In contrast, genes of the EDNRB pathway contributed to only 8 cases (4.2%).

Frequency of Copy-Number Variants in Hirschsprung's Disease:

Of the 190 patients, 17 (8.9%) had syndromic presentations or known major chromosomal variants (Table 4). To detect subkaryotypic changes, we examined the exome data to identify large copy-number variants. In total, we identified 16 distinct copy-number variants; 14 of these variants (and their loci) were not previously known to be associated with Hirschsprung's disease (Table 4). We assessed the pathogenicity of each variant on the basis of its enrichment in patients or their association with a known developmental disorder to identify 9 chromosomal variants and copy-number variants in 11.4% of patients (21 of 185), with a corresponding frequency of 0.2% (40 of 19,584) in controls, a highly significant effect (odds ratio, 63.07; 95% CI, 36.75 to 108.25; $P = 4.19\times10^{-51}$) (Tables 2 and 4, and Table S9 in the Supplementary Appendix) (Amiel et al., 2008; Badner et al., 1990; Chakravarti & Lyonnet, 2001).

Of the 21 instances of pathogenic chromosomal variants in patients, 18 (86%) were recurrent and 3 were nonrecurrent, and 18 were in patients with syndromic presentations (Table S9 in the Supplementary Appendix). The most frequent (11 variants, 52%) recurrent finding was trisomy 21, but the other 7 occurred at well-known

loci for other genomic disorders. The elevated frequency of trisomy 21 among patients with Hirschsprung's disease (odds ratio, 73.69; 95% CI, 34.97 to 155.29; $P = 1.23\times10^{-29}$) (Table 2) is not surprising, given previous observations (Arnold et al., 2009). However, the 16p11.2del copy-number variant, which is usually associated with autism (Betancur, 2011), is also significantly enriched (odds ratio, 30.03; 95% CI, 9.70 to 92.97; $P = 3.62\times10^{-9}$). Overall, the 9.7% frequency of patients with Hirschsprung's disease who have recurrent chromosomal variants is significantly higher than the expected frequency (odds ratio, 53.30; 95% CI, 30.30 to 93.76; $P = 2.60\times10^{-43}$). These recurrent chromosomal changes are known to be associated with intellectual disability, autism, neurodevelopmental delay, epilepsy, and Charcot–Marie–Tooth disease type 1A (Betancur, 2011; Lupski et al., 1991), perhaps owing to pathways common to the enteric and central nervous systems. The three nonrecurrent variants, one of which deletes *EDNRB*, were unique, and all occurred in patients with syndromic presentations (Table 4).

Distribution of Diverse Pathogenic Alleles:

Pathogenic alleles in at least 32 genes and loci contribute to Hirschsprung's disease: rare coding variants in 24 genes, common noncoding variants at four sites within 2 loci, and large copy-number variants and chromosomal anomalies in at least 8 additional loci (not including 13q21.33-q31.1del, which overlaps *EDNRB*). The common noncoding risk genotypes (five or more risk alleles), rare coding variants, and copy-number variants occur at decreasing (by orders of magnitude) frequencies in the general

population — 17.1%, 5.0%, and 0.2% —but with increasing odds ratios of 4.54, 10.02, and 63.07, respectively (Table 2). In consequence, all three variant classes make major contributions to the risk of Hirschsprung's disease, with population attributable risks of 37.7%, 31.1%, and 11.3%, respectively, and a total attributable fraction of 61.9%. In addition, although the differences are not significant, the odds ratios among males are consistently higher than those among females (Table S10 in the Supplementary Appendix). Thus, the sex effect in Hirschsprung's disease is not caused by a specific gene or variant but is a property of the disorder. We conclude that, first, even in this rare disorder, common variants are responsible for the majority of cases of Hirschsprung's disease, despite their individually lower risks, because of their high population prevalence. Second, the total risk from all rare coding pathogenic alleles (which have a much higher penetrance) is also high but is differentially spread over 24 genes. Third, the population risk from copy-number variants is the lowest, spread over the effects of 9 loci but with a majority contribution from trisomy 21. These risks from both known and novel genes and loci are almost certainly overestimates owing to the "winner's curse." Consequently, we reestimated the risks, taking into consideration only the well-established risk factors and genes known before this study, and we found the same pattern: these variant classes occur at frequencies of 17.1%, 3.9%, and 0.1% in the general population, but with increasing risks — odds ratios of 4.54, 6.70, and 73.69, respectively (Table 2). These three categories contribute to the population attributable risks of 37.7%, 18.2%, and 9.1%, respectively, or a total attributable fraction of 53.7%.

Finally, we quantified the risk associated with combinations of pathogenic alleles (Table S11 in the Supplementary Appendix) (Angrist et al., 1996). We classified each patient's total burden of pathogenic alleles according to sex, segment length, familiality, and the presence or absence of additional anomalies; we pooled all patients with copy-number variants into one class, given the low frequency of this type of variant. The results showed three cardinal features (Table 5). First, genetic risk factors of any type were identifiable in 72.1% of patients, and patients harbored various combinations of different types of pathogenic alleles, all in significant excess relative to controls. Second, each of the three variant classes (five or more common noncoding variants, rare coding variants, and copy-number variants) were present in substantial percentages of diagnoses (48.4%, 34.7%, and 11.4%, respectively) (Table 2). One, two, or three different classes of molecular lesion were present in 51.9%, 18.4%, and 1.7% of patients, respectively — roughly their expected frequencies — with no evidence of interaction, a finding consistent with multifactorial expectations, although the statistical power for such detection is probably low (Table 5, and Table S11 in the Supplementary Appendix). Third, the genotype-specific odds ratios for Hirschsprung's disease, estimated in reference to the class with no identifiable genetic risk factor, vary by a factor of 67 and increase with the pathogenic allele burden. These data allow us to estimate the absolute risk of Hirschsprung's disease, given a person's genotype. Persons with no identifiable risk factors have an estimated population risk of 5.33 per 100,000 (approximately 1 per 18,800), a low risk of disease. At the other extreme, persons with both common enhancer risk genotypes and rare coding variants and those with copy-

number variants have substantial estimated risks of 2.85 per 1000 (approximately 1 per 350) and 8.38 per 1000 (approximately 1 per 120), respectively.

We did not detect any significant genotype–phenotype associations with respect to sex, segment length, familiality, or syndromic status. However, patients with a copy-number variant and patients with both a common transcription-enhancer risk genotype and a rare coding variant — the two classes with the highest relative risks — are characterized by an excess representation of males and of nonfamilial cases. The sex ratio in classes with no evident pathogenic alleles or those with rare coding single-nucleotide variants only is approximately 1. This latter class is most often seen in persons with an affected relative (familial disease), which suggests that most segregating pathogenic alleles in affected families are rare coding variants. There was also a greater tendency for Hirschsprung's disease to be syndromic among patients in higher risk classes than among those in lower risk classes.

3.4 Discussion

Hirschsprung's disease can arise both from low-penetrance genetic disorders (Badner et al., 1990; Chakravarti & Lyonnet, 2001; Edery et al., 1994; Puffenberger et al., 1994) and from high-penetrance monogenic syndromes (Amiel et al., 2008; Chakravarti & Lyonnet, 2001). Risk prediction and genetic counseling therefore depend on family history, risk factors (sex and segment length), and targeted assessment for syndromic features (Badner et al., 1990). Thus, in a small subset of patients, classical genetic testing of *RET*, *EDNRB*, and genes that are associated with syndromes may be

informative. The results reported here, however, suggest that widespread genomic analyses may be useful for clinical research and improved risk stratification.

Hirschsprung's disease is usually an isolated condition and unassociated with family history. However, genetic causal factors can be identified in approximately 72% of cases, for which molecular class, frequency, and disease risk can be quantified on the basis of sequence data alone, explaining between 53.7% and 61.9% of population attributable risk. Approximately 21% of patients have multiple risk factors, with the genotype-specific incidence increasing by a factor of more than 100 (risk ranging from approximately 1 in 18,800 to 1 in 120) as the number of genotypic risk factors increases from zero to three. Therefore, we have sufficient quantification of disease risk according to genotype to address questions of underlying causes and genetic architecture and to provide genetic counseling for the highest-risk 21% of patients and their relatives.

We have made considerable strides in understanding the functional basis of Hirschsprung's disease. The majority of the 32 genes and loci are known to have roles in the development of the enteric nervous system. In contrast, the majority of patients (63.2%) have identifiable pathogenic alleles only within the known RET regulatory network, which lead to decreased RET signaling. The RET effect is potentially even larger, affecting 78.9% of patients, because an additional 5.8% of patients harbor pathogenic alleles in *UBR4*, a novel E3 ligase gene identified in this study and a candidate for RET signal termination; 5.9% of patients have trisomy 21, which results in an elevated dosage of *SOD1*, encoding a negative regulator of RET (Arnold et al., 2009); and 4.2% of cases involve *EDNRB*, which is SOX10-regulated (Carrasquillo et al., 2002;

Stanchina et al., 2006). Thus, genetic testing of at least the RET regulatory network is warranted for risk stratification.

In order to understand the biology of enteric nervous system cell proliferation, migration, colonization, and neuronal specialization, it is important to understand the steps subsequent to the transition and differentiation of enteric nervous system cells, such as the likely axonal guidance functions of *SEMA3C* and *SEMA3D* (Jiang et al., 2015). The seven novel genes identified here, all of which are expressed in the human gut at the appropriate developmental stages, probably control some aspects of axonal guidance, cell proliferation, and local inflammation (Table S12 in the Supplementary Appendix). We hypothesize that screening the genes regulating these processes in early gut development will further resolve the remaining approximately 40% of Hirschsprung's disease risk.

A continuing challenge in the study of Hirschsprung's disease is to understand the cellular mechanisms underlying the disease. Whether we consider the persons with the highest (1 in 120) or lowest (1 in 18,800) risk, the absolute risk of disease is still small. What are the cellular events that trigger or prevent aganglionosis, given a particular genotype? A part of the answer is the existence of very rare de novo gene mutations affecting the enteric nervous system (Gui et al., 2017), which require larger cohorts of trios for the detection of an association. However, the incomplete penetrance of most Hirschsprung's disease variants implies that stochastic, environmental, or epigenetic factors must be important.

In our study, we found that the risk of the complex phenotype that is Hirschsprung's disease stemmed from a combination of variants in numerous genes and different classes of genetic variants: noncoding variants, single nucleotide variants and copy-number variants, and both rare and common variants. Despite the current thinking in human medical genetics, most of the risk of Hirschsprung's disease arose from a common widespread genetic susceptibility, on top of which rare coding and rarer copy-number variants exacerbated the risk. Despite this molecular diversity, the implicated genes clustered, on the basis of their known function, into gene regulatory networks (which, in Hirschsprung's disease, regulate the transition from enteric neural crest cells to enteric neuroblasts, axonal guidance, and neuroblast proliferation), a model that may be relevant to the understanding of other complex disorders.

3.5 Support and Acknowledgements

Supported by

Supported by grants from the National Institutes of Health (MERIT award R37 HD28088 to Dr. Chakravarti, R01 MH101221 to Dr. Eichler, and U54 HG003067 to Dr. Gabriel). Dr. Eichler is a Howard Hughes Medical Institute investigator.

Financial disclosure

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

Acknowledgments

We thank the numerous patients and their family members, physicians, and genetic counselors who have contributed to these studies over the years; and Erick Kaufmann, Jennifer (Scott) Bubb, Sue Lewis, Maura Kenton, and Julie Albertus for family ascertainment and genetic counseling.

3.6 Chapter 3 Tables:

Table 1. *Population risk of Hirschsprung's disease as a function of RET and SEMA3 non-coding risk allele dosage*.

# risk alleles	Number (%) of cases (n = 186)	Number (%) of controls (n = 627)	Odds ratio (95% CI)	*P*
1	9 (4.8)	87 (13.9)	(1)	-
2	13 (7.0)	137 (21.9)	0.90 (0.38-2.13)	8.24×10^{-1}
3	40 (21.5)	172 (27.4)	2.16 (1.03-4.52)	4.28×10^{-2}
4	34 (18.3)	124 (19.8)	2.55 (1.20-5.42)	1.51×10^{-2}
5	35 (18.8)	84 (13.4)	**3.87** (1.81-8.29)	3.01×10^{-4}
6	41 (22.0)	18 (2.9)	**20.66** (8.84-48.31)	6.03×10^{-15}
7 & 8	14 (7.5)	5 (0.8)	**24.28** (7.68-76.74)	1.72×10^{-8}

A total of 186 patients and 627 controls (404 samples from persons of European ancestry [excluding Finns] from the 1000 Genomes Project and 223 pseudo-controls (generated from the chromosomes not transmitted to the affected child from 254 parent–child trios with complete genotype data) were classified according to the number of Hirschsprung's disease risk alleles present at *RET* single-nucleotide polymorphisms (SNPs) rs2435357, rs7069590, and rs2506030 and *SEMA3* SNP rs11766001. All patients had at least one risk allele. Bold odds ratios indicated a significant association (calculated with a two-sided Fisher's exact test) at a family-wise error rate of 0.05, after correction for performing six tests.

Table 2. *Distribution of Hirschsprung's disease risk by the molecular class of risk alleles*.

Disease-associated risk allele class		# independent genes/loci	% frequency		Odds ratio (95% CI)[a]	% Population attributable risk[b]
			cases	controls		
Enhancers, common variants[c]	known	2	48.4	17.1	4.54 (3.19 – 6.46) ($P = 1.22 \times 10^{-16}$)	37.7
Coding genes, rare variants[d]	known & novel	24	34.7	5.0	10.02 (6.45 – 15.58) ($P = 3.41 \times 10^{-25}$)	31.1
	known	17	21.6	3.9	6.70 (4.06 – 11.04) ($P = 9.65 \times 10^{-14}$)	18.2
Copy number changes, rare variants[e]	known & novel	9	11.4	0.20	63.07 (36.75 – 108.25) ($P = 4.19 \times 10^{-51}$)	11.3
	Known[f]	1	5.9	0.09	73.69 (34.97 – 155.29) ($P = 1.23 \times 10^{-29}$)	9.1

[a]Although the differences were not significant, the odds ratios in males were consistently larger than those in females (Table S10 in the Supplementary Appendix).

[b] The combined population attributable risk for all three classes of pathogenic alleles, under the assumption of independent effects, is 61.9% for all 24 known and novel loci and 53.7% for the 18 known loci.

[c] Five or more common disease variants were observed in 90 of 186 patients and 107 of 627 controls.

[d]Rare coding sequence variants were identified in 66 of 190 patients and in 37 of 740 controls.

[e]The copy-number variants (CNV) that were considered to be pathogenic, as reported in this table and in all our other analyses of risk, were clinically identified alterations (e.g., trisomy 21 or 22q deletion) or deletions of more than 500 kb or duplication of more than 1000 kb, with a frequency of less than 1% among controls, that had previously been significantly associated with a developmental disorder. CNVs were identified in 21 of 185 patients and in 40 of 19,584 controls.

[f] The only copy-number variant that was previously known to be associated with Hirschsprung's disease was trisomy 21.

Table 3. *Genes with an excess of rare coding pathogenic alleles in Hirschsprung's disease*.

Gene[a]	# distinct PAs in 190 cases		*P*	Embryonic intestinal gene expression		Zebrafish cell migratory defect[b]
	Observed	Expected		Human (CS22)	Mouse (E10.5)	
EDNRB[*]	6	0.22	1.35×10^{-7}	Yes	Yes	Yes
ADAMTS17[*]	5	0.23	4.45×10^{-6}	Yes	Yes	NT
ACSS2[*]	6	0.45	8.08×10^{-6}	Yes	Yes	Yes
RET	7	0.99	7.73×10^{-5}	Yes	Yes	Yes
SLC27A4	4	0.22	8.48×10^{-5}	Yes	Yes	No
SH3PXD2A	4	0.22	8.88×10^{-5}	Yes	Yes	Yes
MMAA	4	0.23	9.26×10^{-5}	No	No	No
ENO3	5	0.45	1.03×10^{-4}	Yes	Yes	Yes
FAM213A	4	0.40	7.51×10^{-4}	Yes	Yes	No
UBR4	11	3.47	9.52×10^{-4}	Yes	Yes	Yes

[a]Genes that are statistically significant after multiple test correction for 4,027 genes are marked by an asterisk.

[b] NT: not tested owing to the lack of an identifiable zebrafish ortholog.

Table 4. *Karyotypes and large copy number variants (CNVs) in Hirschsprung's disease*.

Karyotype/copy number variant[a]	Size (kb)[b]	Syndrome present	Detection method[c]	# CNVs[d]		*P*[e]
				cases (n =185)	controls (n = 19,584)	
Recurrent Variants						
Free & mosaic trisomy 21*	47,710	Yes	K (9 of 11)	11	17[#]	**6.68×10^{-16}**
16p11.2 del*	740 - 985	1 Yes/2 No	K (1 of 3), E (3 of 3)	3	12	**3.38×10^{-4}**
1q21.1 dup	509 – 1,185	No	E (3 of 3)	3	27	**2.72×10^{-3}**
1q21.1 del*	1,425	Yes	E	1	6	6.37×10^{-2}
22q11.2 del*	8,000	Yes	K	1	0	9.36×10^{-3}
Tetrasomy 22q*	1,447	Yes	K	1	0	9.36×10^{-3}
17p11.2 dup (*CMT1A*) *	1,835	No	E	1	5	5.49×10^{-2}
Non-Recurrent Variants						
47, XX, +der (15) t(4:15)*	7,768 (chr4), 3,800 (chr 15)	Yes	K, E	1	0	9.36×10^{-3}
1p33 del	582	No	E	1	0	9.36×10^{-3}
12p13.31 del	554	Yes	E	1	0	9.36×10^{-3}
13q21.33-q31.1 del*	14,356	Yes	K	1	0	9.36×10^{-3}
2q21.2-q22.2 del*	8,847	Yes	E	1	0	9.36×10^{-3}
8p23.3 del	579	Yes	E	1	0	9.36×10^{-3}
2p25.3 dup	1,377	No	E	1	1	1.86×10^{-2}
7q21.12 dup	1,498	No	E	1	11	1.06×10^{-1}
10q24.3-q26.13 inv	25,600	Yes	K	1	-	-

[a] Karyotype or CNV locus defined by genomic coordinates with prior evidence of pathogenicity marked by an asterisk; [b] Estimated smallest region from karyotype, exome sequence or SNP array data; [c] Detection by karyotyping (K) and exome sequencing (E), with validation by SNP array including two trisomy 21 cases on whom we did not have a submitted karyotype; [d] Controls are from reference 21 but counts for 47, XX, +der(15) t(4:15) are for the two duplications at the translocation site, not the translocation, while control counts were not available and not expected for the 10q24.3-q26.13 inversion; [e] Bold P values are significant after multiple test correction; [#] Estimated from population data (**Table S8**).

Table 5. *Distribution of Hirschsprung's disease cases by genetic risk profile and population effects*.

Risk class (+/-, present/absent)			# cases		Odds ratio (95% CI)[f]	Estimated population incidence[g]	#(%) Male	#(%) Short Segment	#(%) Simplex	#(%) non-syndromic
Non-coding[a]	Coding[b]	CNV[c]	Obs.[d]	Exp.[e]						
-	-	-	50 (28)	140.69	(1)	5.33×10^{-5}	26 (52)	17 (46)	28 (56)	38 (76)
+	-	-	53 (30)	29.02	5.07 (2.93 – 8.76)	2.74×10^{-4}	35 (66)	24 (57)	36 (68)	46 (87)
-	+	-	27 (15)	7.41	9.73 (4.22 – 22.39)	5.47×10^{-4}	14 (52)	9 (41)	14 (52)	17 (63)
+	+	-	29 (16)	1.53	40.68 (10.84 – 152.67)	2.85×10^{-3}	24 (83)	14 (58)	24 (83)	20 (69)
-	-	+	20 (11)	0.36	66.80 (11.44 – 389.96)	8.38×10^{-3}	29 (81)	18 (58)	30 (83)	21 (58)
+	-	+								
-	+	+								
+	+	+								

[a] Common variant: 5 or more risk alleles at *RET* (rs2435357, rs2506030, rs7069590) and *SEMA3D* (rs11766001); [b] Coding: at least one rare, deleterious variant in any of the 24 HSCR susceptibility genes; [c] CNV (copy number variant): A clinically identified alteration (trisomy 21, 22q deletion, etc.) or rare deletion >500kb or duplication >1000kb that had previously been associated with a developmental disorder; [d] Numbers (%) of 179 cases with complete data for all three mutation classes; [e] Expected values are calculated from control frequencies in Table 2; [f] Odds ratios are calculated based on HSCR subjects (reference) having no detectable disease associated variants as defined in a-c; [g] Population incidences are calculated by assuming a total rate of 15 HSCR cases per 100,000 live births. Note that the observed numbers (50, 53, 27, 29 and 20) of the variant classes are close to expected numbers estimated from random association of disease variants (53.44, 50.12, 28.40, 26.64 and 20.41; $\chi^2 = 0.67$, 1df, P = 0.41).

3.7 Chapter 3 Supplementary Appendix

3.7.1 Supplementary Methods

Author contributions:

J.M.T, T.M.T., S.C., A.K., C.B. and A.C. designed experiments. J.M.T, C.B. and A.C. wrote and edited the initial draft of the manuscript and contributed to genotypic risk estimation. C.B. and A.C. selected HSCR families for inclusion in this study. J.M.T., A.Y.L., T.N.T., K.H.N. and C.B. contributed to calling and analysis of coding variation. N.K. performed exome-based CNV calling. A.Y.L. and T.N.T. contributed to CNV validation. M.X.S. performed zebrafish morpholino assays and analysis. S.C. performed gene expression assays and analysis. A.K. performed common risk polymorphism genotyping and analysis. B.C. performed control CNV genotyping. B.C. and E.E.E. contributed to functional classification and analysis of large CNVs. N.G. and S.G. performed genomics data generation. All authors had the opportunity to comment on and approve of the final manuscript.

Sample ascertainment:

Affected individuals were selected from our collection of 636 families comprising their phenotypes, medical, pathologic and family history, and a blood/cell line/DNA sample. Affected persons were classified by segment length of aganglionosis into three groups: short-segment (S-HSCR: aganglionosis up to the upper splenic flexure), long-segment (L-HSCR: aganglionosis beyond the splenic flexure) and total colonic aganglionosis (TCA). In addition, they were also classified by gender, familiality (positive

family history) and occurrence of anomalies other than aganglionosis. We chose a sample of 304 HSCR cases for exome sequencing based on DNA availability and consent for genome studies. For sequence analyses, after data cleaning and quality control, we retained 190 independent, unrelated affecteds and their 47 affected relatives (data version 1.3); for this study, we did not use data on 35 individuals from a genetically isolated Old Order Mennonite population (Puffenberger et al., 1994), 5 samples with poor sequence quality, 24 admixed individuals and 3 individuals whose genetic relationships could not be verified against their pedigrees. The 190 independent unrelated individuals, whom we designate as 'probands,' were most often the actual proband but rarely an affected first degree relative with more complete data. The included individuals self-identified as being of European ancestry, which was checked for consistency with their genotype data (Figure S2). The case sample was composed of: (1) 122 (64%) males and 68 (36%) females; (2) 82 (43%) S-HSCR, 67 (35%) L-HSCR/TCA and 41 (22%) unknown (unspecified) segment length; (3) 125 (66%) simplex and 65 (34%) multiplex families (24 sibs, 20 parent-child, 21 greater than first-degree); (4) 130 (68%) non-syndromic, 6 (3%) single gene syndromes (3 Central Congenital Hypoventilation syndrome (CCHS) and 1 each of Waardenburg (WS), L1CAM (L1CAMS) and Bardet-Biedl (BBS) syndromes), 17 (9%) chromosomal variants (11 with Down syndrome and 1 each with 16p11.2 del, 22q11.2 del, tetrasomy 22q, 47, XX, +der(15) t(4:15), 13q21.33-q31.1 del, 10q24.3-q26.13 inv) and 37 (20%) with multiple anomalies not recognized as a specific syndrome. This sample selection had features comparable to our total collection of 636 probands, except for an oversampling of known segment

length cases, and comprised: 67%, 33% male/female; 39%, 29%, 32% S-/L-&TCA/unspecified; 70%, 30% simplex/multiplex; and, 63%/37% non-syndromic/syndromic. Finally, the sampled sibship size was 1, 2, 3 or 4 for 155, 23, 7 and 1 individual, respectively. Subject ascertainment was conducted with written informed consent approved by the Institutional Review Board of Johns Hopkins University School of Medicine.

For European ancestry controls, we used publicly available exome sequence data from 370 NIMH controls (https://www.nimhgenetics.org/available_data/controls/) and 370 EUR 1000 Genome (1000G henceforth) samples (85 Toscani in Italy, 97 Utah residents of Northern or Western European ancestry, 96 Iberians in Spain, and 92 British in England and Scotland, but excluding 101 Finns owing to possible founder effects) (www.1000genomes.org) (The 1000 Genomes Project Consortium, 2015), generated using the same reagents and procedures by the Broad Institute. For assessing admixture, we included all 2,302 1000G individuals with diverse ancestries.

For common non-coding variant studies, we used a different set of controls genotyped in our laboratory because some of these genotypes were not publicly available: 404 EUR 1000G samples (excluding Finns) and an additional 223 pseudo-controls, generated from the chromosomes not transmitted to the affected from 254 HSCR parent-child trios (Kapoor et al., 2015). The differing numbers of EUR 1000G samples used depended on when the data were accessed and the numbers of samples available at that time.

For copy number variant (CNV) analyses, we used a third control set of 19,584

adult subjects of European ancestry (Coe et al., 2014). Different European ancestry controls were required to accommodate risk factors of different frequencies and the assays available in control samples.

A summary of all cohorts studied and of the assays performed on them can be found in **Supplementary Figure S9**.

Genotyping:

Genotype data for the polymorphisms rs2435357, rs2506030 and rs7069590 at *RET* and rs11766001 at *SEMA3C/D* were previously generated using Taqman assays in our laboratory, and have been previously reported (Chatterjee et al., 2016; Kapoor et al., 2015). In addition, HSCR cases with large copy number variants (CNVs; see below), together with their parents where available, were validated by genotyping using the Human Omni 2.5-4v1 BeadChip, using standard methods at the Broad Institute.

Exome sequencing, variant calling and annotation:

Genomic DNA was used to capture exomes using the Agilent 44Mb Sure-Select Human All Exon v2.0 capture, and sequenced using the 76 base paired-end method on an Illumina HiSeq2000 sequencer with >80% of bases at a coverage of 30X. The sequence data were aligned by the Picard (http://picard.sourceforge.net) pipeline using hg19 with the BWA algorithm and processed with the Genome Analysis Toolkit (GATK) (DePristo et al., 2011) to recalibrate base-quality scores and perform local realignment around known insertions and deletions (INDELs) (H. Li & Durbin, 2009). BAM files were used to call single nucleotide variants (SNVs) and small (<50bp) insertions and deletions (INDELs) using the GATK's HaplotypeCaller algorithm. Variants were called

simultaneously across all HSCR cohort members and controls (amounting to 3,176 samples) into a single VCF file. Initial filtering was done via the Variant Quality Score Recalibration (VQSR) method within GATK, which is based on detection of known variant sites. For SNVs, HapMap3.3 and Omni2.5 were used as training sites with HapMap3.3 used as the truth set. SNVs were filtered to obtain the highest confidence variant set achieving 99% truth sensitivity (1% false negative rate). For VQSR of INDELs, a set of curated INDELs obtained from the GATK resource bundle (Mills_and_1000G_gold_standard.indels.b37.vcf) were used as both a training and truth set. INDELs were filtered to obtain the highest confidence variant set achieving 91% truth sensitivity (9% false negative rate). Following initial filtering, an additional annotation was added for ancestral alleles using: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/human_ancestor_GRCh37_e59.README.

SNVs and INDELs meeting initial filtering criteria were further filtered using several hard quality filters. First, all multi-allelic sites were removed. Second, they were filtered on strand balance and homopolymer criteria (FS > 50 and HRun > 5.0 for SNVs, FS > 200 and HRun > 10.0 for INDELs). Third, all individual genotypes with a depth < 10 were removed. Lastly, the dataset was filtered by deleting variants with >10% missing genotypes, separately for autosomes and sex chromosomes, and separately for males and females. ANNOVAR was used for annotation of variants (Wang et al., 2010). Average 46-way PhyloP conservation scores were added to each variant using internal lab scripts.

We assessed whether coding variant coverage and detection sensitivity were comparable between the 190 cases and 740 controls by summarizing sequencing coverage of coding genes targeted by our exome capture reagents (and the 24 HSCR genes specifically) and by counting singleton and doubleton variant sites per individual in each set, across all genes. As shown in **Figure S8**, these metrics are comparable; there are no significant differences across case and control exomes, except that the variance in coverage is smaller in cases. Thus, the sensitivity of variation detection is identical between cases and controls excluding the possibility of false associations through differences in sensitivity.

CNV analysis using exome sequence data:

Mapping, CoNIFER and CNV Segmentation: Short reads from the exome sequencing experiment were split into 36bp chunks and mapped using the single-end mode of mrsFAST (up to two mismatches) to exons and 300bp flanking sequence extracted from the repeat-masked hg19 reference genome, using the target file for the Agilent SureSelect Target Enrichment capture platform. Next, we used CoNIFER and calculated RPKM values for 189,894 probes and exons derived from the target file. We set the --svd option to 12, and used default CoNIFER settings for all other options. Subsequently, the raw SVD-ZRPKM values were exported for downstream analysis. We used DNACopy and CGHCall to segment and assign probabilities to SVD-ZRPKM values. To prevent excessively strong SVD-ZRPKM signals from interfering with the models used by CGHCall to assign copy number, we clipped the signal at ±3 for each exon. Parameters for DNACopy were set to default and the alpha parameter was set to 0.01.

Default options for CGHcall were used, and we allowed only “deletion” and “duplication” as called states. Using these parameters, we obtained 13,300 raw segments “deleted” or “duplicated”. These computational methods are not optimized to detect aneuploidies because the data are normalized by chromosome within each sample.

Quality Control: We excluded samples with more than 200 calls after segmentation, as such sample have extremely high false discovery rates (FDR). Four samples (HSCR274, HSCR18, HSCR385 and HSCR178) with a total of 1,939 calls were excluded, leaving 11,361 calls.

CNVR generation and call filtering: We clustered calls using a custom hierarchical clustering method which uses the pairwise reciprocal overlap (RO) between calls as a measure of distance. To prevent merging of large unique calls, the RO function was modified by a gamma tuning parameter, which weights the RO based on the total number of non-overlapping probes on each end. In this way, the function accounts for the uncertainty in breakpoints and RO for two small CNVs, while allowing two large overlapping CNVs to be counted as distinct entities. Calls were merged using hierarchical clustering (WPGMA, weighted pair group method with averaging), and we flattened the resulting trees to form CNVR clusters. Using this method, we generated 3,129 CNVRs.

Filtering segmentally duplicated regions and processed pseudogenes: We excluded CNVs which were found to have more than 50% of their probes within segmental duplications or duplicated regions of the genome (defined using previous

methods from 1000 Genomes whole-genome depth-of-coverage analysis, where >80% of 34 unrelated genomes had a copy number three or greater in 500bp repeat-masked windows across the genome). Excluding calls which overlapped at least 50% with these regions resulted in the exclusion of 5,079 calls (45% of all calls), corresponding to 525 CNVRs. Next, we excluded calls which were likely to be due solely to the insertion of processed pseudogenes. CoNIFER, and most exome-based read-depth methods, are sensitive to copy number changes, specifically of exons, which can be the result of retro-insertion of processed mRNA transcripts (Krumm et al., 2012). We used two lists of commonly polymorphic processed pseudogenes generated using SPLIT-READ from 225 autism trios (data not reported here) (Karakoc et al., 2012). We excluded calls from our call list for which ≥90% of the probes corresponded to a gene which had been observed at least twice in 225 trios. This excluded a total of 1,063 of 10,927 calls. In sum, our processed pseudo-genes, segmental duplications and other duplicated portions of the genome accounted for 5,808 calls in 673 CNVRs.

Final filtering and call set generation: Our final set of calls and CNVRs was created by requiring at least one call in a CNVR with the following attributes: 1) an absolute median SVD-ZRPKM score (i.e., signal strength) of ≥ 1.5, 2) a CGHCall posterior probability of 0.95 or greater, and 3) passing additional filters for duplicated genes and regions as described above. Our final high-quality set of CNVRs contains 1,597 calls in 554 CNVRs. For the current study, we restricted attention to only 111 rare large CNVs, defined as deletions >500kb and duplications >1mb, and potentially with phenotypic impact (as assessed using external databases), but we also included all changes detected

by karyotype or FISH (fluorescent *in situ* hybridization) for clinical diagnosis, loci for known genomic disorders and HSCR genes. All these cases were further validated by SNP genotyping using the Human Omni 2.5-4v1 BeadChip array (see **Genotyping**).

Validating CNV calls: Omni2.5 Beadchip genotype data were processed using a standard Illumina pipeline; 4 samples failed the QC process. We manually examined the data to confirm each CNV by plotting B allele frequency and the LogR Ratio (LRR) for each from the gtc file. For each chromosomal position, we ignored samples if either the B allele frequency or LogR Ratio was missing, if the GCScore ≤ 0.15, and if multiple discordant genotype calls were made. We also noted the following: annotated genes in the region, and exome sequencing coverage. We validated 31 cases, as shown in **Table 3** and **Supplementary Table S8**, 22 cases of 8 unique recurrent CNVs and 9 non-recurrent CNVs. Of the 14 CNVs not evident from karyotypes, 13 had the median SVD-ZRPKM cutoff >1.5, and 1 between 1 and 1.5. Of the 17 validated CNVs in 31 cases, we could determine parental origin in 8 cases: 4 were *de novo* (del13, +21, 1q21.1 del, inv 10), 2 were inherited from the father (17p11.2 (*CMT1A*) dup, dup7), 1 was inherited from the mother (t4;15); 1 was not inherited from the father but maternal origin could not be assessed since her sample failed QC (+21). We also identified an additional 4 kb *RET* deletion (chr10:42917793-42922026) in patient HSCR472 which was separately validated by qPCR analysis; note that the chromosome 13q21.33-q31.1 deletion included *EDNRB*.

Statistical Analyses:

Principal component analysis (PCA): To assess population structure and potential cases of admixture, we conducted standard PCA analysis using R package SNPRelate (Zheng et al., 2012) on all 301 HSCR cases and 2,672 controls (370 NIMH samples and 2,302 1000G) to identify 29 highly-admixed HSCR individuals (potentially African- and Asian- Americans). We used genotypes for 7,536 autosomal SNPs from exome sequencing that had allele frequencies ≥10%, missingness <5%, LD trimmed using an r^2 threshold of 0.2. In **Supplementary Figure S2** we show European ancestry HSCR probands and all 2,672 PCA controls plotted along the first three PCs, followed by the first three PC's of a Europeans-only PCA, showing that the only European ancestry group from whom HSCR probands can be discriminated is Finns.

Sequence similarity between relatives: The exome sequence data were used to assess the overall genetic relationship between each case and his/her relative. Our sample included 42 relatives of 190 probands yielding 53 relative pairs from 32 families. We used the exome sequences of each pair to compute a similarity statistic S:

$$S = \frac{n_{xy}\left(1/n_x + 1/n_y\right)}{2}$$

where, n_x, n_y and n_{xy} refer to the number of distinct alleles at variant sites in individual *x*, in individual *y*, and shared by *x* and *y*, respectively, at a variant site and is summed across all variant sites (C. C. Li et al., 1993). $S = 0$ whenever n_x, n_y or n_{xy} is zero. *S* is the proportion of shared sites relative to the harmonic mean of the number of variants in the pair compared. The coefficient of relationship (*r*) is then estimated as:

$$r = (\bar{S} - U)/(1 - U),$$

where $\bar{S}$ is the average *S* across all variants, estimated by summing the numerator and denominator in the above formula, and *U* is the similarity statistic from unrelated individuals, estimated from all possible pairings of the 190 cases (C. C. Li et al., 1993).

To assess whether susceptibility variants were enriched in affected relatives of probands, we estimated *S* for the 24 HSCR genes, based on pathogenic alleles. We estimated the mean *S* for all such relative pairs and compared it to its expectation by obtaining its empirical distribution from 5,000 means based on 24 randomly selected genes for each relative pair, restricting analysis to only those gene sets with at least one pathogenic allele in the proband. This distribution was used to calculate a one-sided test of excess sharing in these relatives. Across all relative pairs, $\bar{S}$ = 0.75 and was significantly greater than the mean permuted value of 0.45 (*P* = 0.0054).

Discovery of genes enriched for rare coding single nucleotide variants (SNV): We compared rare pathogenic SNVs, defined as coding alleles that are nonsense, highly conserved missense (PhyloP score ≥ 4) or changes that alter the canonical ±2bp splice junctions, with frequency $\leq$ 5% in cases and controls for genes in which at least one pathogenic SNV was observed in controls. For analysis, we used the observed number (*d*) of distinct pathogenic SNVs among the 190 cases (d_o), motivated by the known distribution of allele multiplicity for alleles of a defined selection coefficient (Hartl & Campbell, 1982). To assess whether the observed value was higher than expected we used 740 European ancestry NIMH and 1000G controls to randomly sample 190 individuals and calculate *d* for each replicate. Repeating this sampling 10,000 times, with replacement, provided an estimate of the distribution of the random variable *d* with

mean $\bar{d}$. We estimated the significance value (α) of the hypothesis of no gene effect as α = Prob $\{d \geq d_o | \bar{d}\}$ and by assuming *d* is Poisson distributed, an assumption that was tested from the empirical distribution of *d* across the replicates. This assumption was conservative since the observed variance of the distribution was smaller than the average (**Supplementary Figure S4**). Note that these are gene-specific estimates and so no corrections for gene size or sequence features are necessary, although the statistical power of detecting departures in individual genes decreases with a gene's increasing intrinsic rate of pathogenic variation in controls. This empirical probability distribution, contrasted to the expected distribution, for all human genes was used for testing whether there is an excess of pathogenic variants in specific genes (see QQ plot in **Supplementary Figure S4**).

Discovery of copy number variants (CNV): CNV burden was compared between cases and controls for rare CNVs (prevalence <1%) using CNV length, excluding gaps and regions annotated as segmental duplications (hg18). The 19,584 controls (Coe et al., 2014) were obtained by combining 8,329 controls from Cooper *et al*. (dbVar study accession nsdt54) with 11,255 new controls profiled on Affymetrix SNP6 arrays from the Wellcome Trust Case Control Consortium 2 (WTCCC2) 58C cohort, as well as the ARIC (Atherosclerosis Risk in Communities) Cohort (database of Genotypes and Phenotypes, dbGaP accession phs000090.v1.p1) (Supplementary Table 1 in (Coe et al., 2014). The details of CNV calling in controls are described there. CNV calls that falsely extended across centromeric gaps, due to small polymorphisms on both arms, were trimmed. These CNVs are shown in Supplementary Figure 1 of Coe et al., 2014. Burden was

defined using only the largest CNV to account for the large number of bases encompassed by small CNVs and the difference in resolution between cases (exome sequence) and controls (SNP arrays). The overall incidence of rare deletions and duplications among these 19,584 controls was 0.020 (391 instances) and 0.014 (282 instances), respectively. These controls did not include individuals with intellectual disability and so this estimate was supplemented by the prevalence of Down syndrome in the population (8.27/10,000 individuals) (Presson et al., 2013), and its value imputed for controls of equivalent size (i.e., 19,584).

Quantifying pathogenic allele (PA) enrichment: We tested for enrichment of PAs by class in individuals of European ancestry. (1) Common variants were allele, haplotype and genotype counted in 186 cases and 627 controls with tests conducted using contingency chi-square methods with significance calculated using a 2-sided Fisher's exact probability. Frequency differences were represented as corrected odds ratios (Kapoor et al., 2015) with variances and tests of significance as estimated using the Haldane bias correction (Haldane, 1956). (2) Rare coding variants were compared using exome sequence data in 190 cases and 740 1000G and NIMH controls. For overall enrichment, the PA definition for rare coding variants was extended to include all INDELs overlapping the coding sequence and restricted to only PAs identified in cases. For these alleles, we report their allele frequencies in ancestry-matched, non-neuropsychiatric samples from ExAC (Exome Aggregation Consortium et al., 2016), comprising a much larger sample size of 21,071 subjects. Frequency differences were represented as corrected (owing to small numbers in some cells) odds ratios (Kapoor et

al., 2015) with variances and tests of significance as estimated using the Haldane bias correction (Haldane, 1956). (3) Each recurrent and non-recurrent CNV was compared against its frequency in a sample of 19,584 adult subjects of European ancestry. Tests of enrichment used a 2-sided Fisher's exact probability. Note that these controls were all ascertained as adults and therefore were depleted for high penetrance CNVs such as trisomy 21, which we corrected for individually. The other CNVs detected are not *a priori* known to lead to high penetrance phenotypes.

Estimating disease penetrance: Phenotype penetrance refers to the probability of phenotypic expression for specific genotypes and, as such, are marginal effects averaging across the phenotypic effects (if any) across all other genes (genetic background). As such, this penetrance is also the disease incidence given that genotype. The incidence is the frequency (rate) of new cases which may manifest or be recognized at different ages. However, for HSCR these are nearly identical because most cases are recognized and treated in the first years of their life. Consequently, genotype-dependent penetrance can be calculated as follows:

$$P\{D|G\} = P\{G|D\}P\{D\}/P\{G\},$$

where G, D and D^c are genotype, phenotype (disease) and the phenotype complement, respectively, and P {.} is probability. If we examined n cases and m controls, and P{G} was the population frequency of a specific genotype class (either at one locus or at many) then it could be estimated from the control data, while P{G|D} could be estimated from the genotype distribution among cases. P{D} is the disease incidence

and for HSCR is set to 15/100,000 live births. The penetrance estimated from the above equation has the expected standard deviation of:

$$P\{D|G\}\sqrt{\frac{1-P\{G|D\}}{nP\{G|D\}}+\frac{1}{mP\{G\}}},$$

which is estimated by replacing all expected values by their observed quantities.

Estimating population attributable risk: Population attributable risk (PAR) is the proportion of disease in a population involving a given exposure, which is useful in this study for comparing the relative contributions of each risk factor to the development of HSCR. In order to estimate PAR, one requires an estimate of the relative risk (RR) of each risk factor as well as the proportion of the general population exposed. When disease incidence is low, as is the case with HSCR, RR ~ OR. Therefore, PAR can be estimated for HSCR risk factors as follows:

$$PAR = \frac{P_e\,(OR-1)}{P_e\,(OR-1)+1}$$

where P_e is the proportion of the general population exposed to that risk factor derived from one of our three control populations.

Gene expression studies:

Taqman gene expression assays of human and mouse gut tissue: We studied 8 human fetal guts at Carnegie Stage 22 (CS22) obtained from the Human Developmental Biology Resource (www.hdbr.org; grant 099175/Z/12/Z). All HDBR samples were collected according to local Research Ethics Committee review by the NHS Health Research Authority National Research Ethics Service and in line with the ethical

guidelines laid out in the Polkinghorne Report (Review of the Guidance on the Research Use of Fetuses and Fetal Material, 1989). The HDBR is also licensed as a tissue bank by the Human Tissue Authority. The samples were approved for use in this study by the Institutional Review Board of Johns Hopkins University School of Medicine. All mouse guts used were from three E10.5 wild type C57BL/6J male mice purchased from The Jackson Laboratory. Total RNA was extracted from these tissues using TRIzol (Life Technologies, USA) and cleaned on RNeasy columns (Qiagen, USA). 500ng of total RNA was converted to cDNA using SuperScriptIII reverse transcriptase (Life Technologies, USA) and Oligo-dT primers. The diluted (1/5) total cDNA was subjected to Taqman gene expression (Life Technologies, USA) using transcript-specific probes and primers. Human or mouse β-actin was used as an internal loading control to normalize data. Each sample was assayed 3 times and the data presented are means with their standard errors. The relative fold-change was calculated based on the $2^{\Delta\Delta Ct}$ (threshold cycle) method, with the highest expressing transcript (lowest Ct value) set to unity. Any gene with Ct value >30 was considered not expressed. Only one potential gene- *MMAA* had a Ct value >30 in both mouse and human. The following Taqman probes were used from Applied Biosystems: For human: *RET* (Hs01120032_m1), *EDNRB (*Hs00240747_m1*), ADAMTS17 (*Hs00330236_m1*), ACSS2* (Hs01120914_m1*), SLC27A4 (*Hs00192700_m1*), SH3PXD2A (*Hs01046313_m1*), MMAA (*Hs00604098_m1*), ENO3 (*Hs01093275_m1*), FAM213A (*Hs00800009_s1*)* and *UBR4 (*Hs00390223_m1*).* For mouse: *Ret* (Mm00436305_m1), *Ednrb* (Mm00432989_m1), *Adamts17* (Mm01318914_m1), *Acss2* (Mm00480101_m1), *Slc27a4* (Mm01327405_m1), *Sh3pxd2a* (Mm01205065_m1), *Mmaa*

(Mm04209905_m1), *Eno3* (Mm00468267_m1), *Fam213a* (Mm00510430_m1) and *Ubr4* (Mm01348737_m1).

RNA-seq gene expression assays of mouse gut tissue: Total RNA was extracted from 3 male mouse guts at E10.5. cDNA was prepared by oligo dT beads to select mRNA from the total RNA sample followed by heat fragmentation and cDNA synthesis from the RNA template as part of the Illumina Tru Seq™ RNA Sample Preparation protocol. The resultant cDNA was used for library preparation (end repair, base 'A' addition, adapter ligation, and enrichment) using standard Illumina protocols. Libraries were sequenced on a HiSeq 2000 using manufacturer's protocols to a depth of 15 million reads per samples (75 base pair, paired end). The primary data were analyzed using the Broad Institute's Picard pipeline, which includes de-multiplexing, and data aggregation. The resultant BAM files were mapped to the mouse genome (assembly mm10/GRCm38) using *TopHat* with its setting for paired end, non-strand specific library. Successfully mapped reads were used to assemble transcripts and estimate their abundances using *Cufflinks* (Trapnell et al., 2012). The resulting data assigned Fragments Per Kilobase of Transcript per Million mapped reads (FPKM) values for each transcript and gene. To further assign which genes were "expressed" in the gut, we did qPCR analysis of multiple genes with FPKM ranging from 1-10. Since we did not always detect expression of genes with FPKM < 5, we set FPKM of 5 as the threshold for genes to be considered gut expressed. All data have been deposited in NCBI's GEO and are accessible at accession number GSE99232**.**

Morpholino studies in zebrafish:

Zebrafish Maintenance and embryo collection: Zebrafish (AB strain) were raised and maintained under standard conditions. All animal research was approved by the Institutional Animal Care and Use Committee at Johns Hopkins University. Embryos were collected and staged as described previously (Kimmel et al., 1995; Westerfield, 1991).

Morpholino microinjections: Translation blocking morpholinos (MO) were designed against each zebrafish ortholog to the human gene and ordered from Gene Tools, LLC along with a standard negative control morpholino; the sequences are provided in Table S13. All genes had a single zebrafish ortholog except *EDNRB* for which both zebrafish orthologs were tested. Injections were performed on 1-2-cell zebrafish embryos (n=50) independently on at least 2 different days. Survival of uninjected, negative control and transcript-specific morpholino-injected embryos were recorded to assess the effect of the transcript-specific morpholinos on survival. Different concentrations were injected for each MO to determine the optimal concentration at which a phenotype was detected. Only two concentrations are being reported for each MO for simplicity; the lowest concentration at which an effect, if any, is seen and the highest concentration before the morpholino is lethal to the embryo.

Immunostaining and visualization: Injected zebrafish embryos were fixed at 6 dpf (days post-fertilization) with 4% paraformaldehyde (PFA). Monoclonal anti-HuC antibody (Invitrogen #A-21271.) followed by Alexa Fluor 568 F $(ab')_2$ fragment of goat anti-mouse IgG secondary antibody (Invitrogen #A11019) were used for fluorescent

labeling of enteric neurons as previously described, with a mild modification (see Jiang et al., 2015) (Kuhlman & Eisen, 2007). The embryos were bleached after fixing in 4% PFA by incubating in 3% H_2O_2/0.5% KOH medium until there was a complete loss of epidermal pigmentation (~30-45 min), followed by a 5 min wash with PBS to stop the bleaching reaction. Stained embryos were visualized using a Nikon SMZ 1500 fluorescent microscope using a DS red filter to assess the colonization of enteric neurons in the gut of each embryo.

Cell counting: Stained neurons were counted using the Image-based Tool for Counting Nuclei (ITCN) plugin in ImageJ visualization software (Abràmoff et al., 2004), with the following parameters: width 9 pixels, minimum distance 4.5 pixels, threshold of 1 and using a selected Region of Interest (ROI). Since the enteric neurons are mostly lost caudally in the gut in well-established HSCR models in zebrafish, we chose our region of interest as 8 somites starting at the caudal end of the gut and going rostral. 15 embryos were used for cell counting for each concentration of morpholino for each gene; 20 embryos were counted for controls.

3.7.2 Chapter 3 Supplementary Tables

Supplementary Table S1: *Genes with disease-associated variants (DAV) and pathogenic alleles (PA) reported in HSCR mutation databases.*

HGMD and ClinVar reported 489 DAVs for HSCR, but our criteria for identifying a PA would have identified a smaller set of 395 (80.8%) alleles. These databases do not specify why most alleles are considered pathogenic. Note that the average allele frequency of these PAs is ~15X smaller than the corresponding DAVs suggesting a greater deleterious effect (penetrance). However, these PAs are a biased set since 66.6% of them are null alleles which are easier to recognize as pathogenic and are, therefore, preferentially reported in the literature: 96 (24.3%) missense, 95 (24.1%) nonsense, 27 (6.8%) splice junction, (42.5%) frame-shifting INDELs, and (2.3%) non-frame-shifting INDELs. As expected, the null alleles are extremely rare and at ~400X lower frequency than all PAs, demonstrating an even greater deleterious effect or higher penetrance. Determining causality for missense variants is much more difficult and requires statistical analysis of enrichment using controls or functional studies or both. Note the wide variation in reported DAVs and PAs across the HSCR genes, including that of null alleles, indicating differential allelic effects across genes. Consequently, the reliance on null alleles only for gene discovery and reporting creates an extreme bias in HSCR, and other genetic studies, for gene identification. Unbiased studies of PAs in different genes require appropriate control data on those same alleles. The overall PA detection rate is not possible to estimate from these data since the total numbers of patients screened were not reported. In contrast, we can estimate the maximum 'false' positive rate of PA detection *under our criteria* at ~13.2% since 52 such PAs were identified in 98 of 740 NIMH and 1000G controls (**Table S6**), in whom knowledge regarding HSCR family history is absent. This is an upper estimate since true causal variants have low penetrance and are expected to appear at low rates in controls.

Gene	*Locus*	*Syndrome*	*# DAVs*[a]	*# PAs*[b]	*# (%) of null PAs*[c]
PHOX2B[1,5]	4p13	Central Congenital Hypoventilation (CCHS)	29	26	22 (84.6%)
SOX10[1,5]	22q13.1	Waardenburg, type 4 (WS4)	38	36	33 (91.7%)
TCF4[1,5]	18q21.2	Pitt Hopkins (PHS)	49	49	32 (65.3%)
ZEB2[1,5]	2q22.3	Mowat Wilson (MWS)	150	144	135 (93.8%)
GDNF[2]	5p13.2	-	5	0	0
NRTN[2]	19p13.3	-	2	0	0
GFRA1[2]	10q25.3	-	2	1	0
RET[2]	10q11.21	-	132	77	29 (37.7%)
ECE1[3]	1p36.12	-	1	0	0
EDN3[3,5]	20q13.32	Waardenburg, type 4 (WS4)	15	6	4 (66.7%)
EDNRB[3,5]	13q22.3	Waardenburg, type 4 (WS4)	42	36	12 (33.3%)
SEMA3C[4]	7q21.11	-	2	2	0
SEMA3D[4]	7q21.11	-	3	2	0
KIF1BP[5] *(KIAA1279)*	1oq22.1	Goldberg Shprintzen (GOSHS)	4	4	3 (75.0%)
L1CAM[5]	Xq28	L1CAM (L1S)	10	8	1 (12.5%)
IKBKAP[5]	9q31.3	Riley-Day (RDS)	2	2	0
NRG1	8p12	-	3	2	1 (50.0%)
Total	-	-	489	395 (80.8%)	263 (66.6%)
Mean allele frequency	-	-	5.53×10^{-4}	3.61×10^{-5}	9.05×10^{-8}

[1] Transcription factors: *PHOX2B, SOX10, TCF4, ZEB2*; [2] RET pathway: *GDNF, NRTN, GFRA1, RET*; [3] EDNRB pathway: *ECE1, EDN3, EDNRB*; [4] SEMA3 pathway: *SEMA3C, SEMA3D*; [5]Single gene syndromes: *PHOX2B, SOX10, TCF4, ZEB2, EDN3, EDNRB, KIF1BP, L1CAM, IKBKAP;* [a] DAV: disease-associated variant as reported in *HGMD* (Stenson et al., 2014) and *ClinVar* (Landrum et al., 2018); [b] PA: pathogenic alleles as defined in Supplementary Methods; [c] Null: nonsense alleles and frame-shifting INDELs. Note that alleles with multiple functional classifications were classified with the following order of priority: nonsense, splice junction, coding INDEL and conserved missense. The mean allele frequency was estimated from non-Finnish European ancestry subjects from the ExAC database; only individuals without a neuro-psychiatric disorder were included (Exome Aggregation Consortium et al., 2016).

Supplementary Table S2: *Four common non-coding variants in Hirschsprung disease.*

190 cases and 627 (404 Non-Finnish EUR 1000G + 223 HSCR pseudo-controls from 254 trios) controls were genotyped for rs2435357, rs2506030 and rs11766001; rs7069590 had genotypes for 186 HSCR samples. The disease associations of these variants have been previously published (Chatterjee et al., 2016; Jiang et al., 2015; Kapoor et al., 2015); their properties in the sample studied here are shown below in **Table S2.1**. Unsurprisingly, the odds ratios have the same magnitudes as reported earlier in larger samples, providing confidence that the sampled cases studied here are representative of HSCR. The genotypes of 186 HSCR cases available for all four markers were next used to count the total number of risk alleles *per individual*, a summary measure of susceptibility arising from common variants in cases and controls, shown below in **Table S2.2**.

Table S2.1: *Common variant associations in HSCR.*

Gene	SNP (risk/non-risk alleles)	Case-control samples*		
		Risk allele (case/control frequency)	Odds ratio (95% CI)	*P*
RET	rs2435357 (T/C)	0.59/0.23	4.8 (3.8-6.1)	6.0×10^{-40}
RET	rs7069590 (T/C)	0.84/0.74	1.8 (1.3-2.5)	9.7×10^{-5}
RET	rs2506030 (G/A)	0.54/0.40	1.7 (1.4-2.2)	2.2×10^{-6}
SEMA3	rs11766001 (C/A)	0.21/0.16	1.4 (1.1-1.9)	0.02

Table S2.2: *Common variant susceptibility distribution in HSCR.*

# risk alleles	Number (%) of cases (n = 186)	Number (%) of controls (n = 627)
0	4 (2.2)	16 (2.6)
1	5 (2.7)	71 (11.3)
2	13 (7.0)	137 (21.9)
3	40 (21.5)	172 (27.4)
4	34 (18.3)	124 (19.8)
5	35 (18.8)	84 (13.4)
6	41 (22.0)	18 (2.9)
7	12 (6.5)	5 (0.8)
8	2 (1.0)	0 (0.0)

Supplementary Table S3*: Exome sequence variation.*

The data used (v1.3) represents joint calling and analysis of 301 HSCR cases and 740 controls all sequenced at the Broad Institute, Cambridge, MA. There were a total of 306,910 SNVs, 3,646 insertions and 7,900 deletions for a total of 318,456 variants. Considering coding sequences only, there were 112,489 SNVs, 844 insertions and 2,753 deletions for a total of 116,086 variants. The properties of these variants passing quality control (**Supplementary Methods**) among the 190 European ancestry cases by type, genomic location and frequency (common defined as a minor allele frequency (MAF) $\geq$ 10%) are as follows:

Variant type	***Coding***			***Total***		
	SNV	***INS/DEL***	***Common***	***SNV***	***INS/DEL***	***Common***
Autosomal	41,295	267/872	6,012	126,089	1,588/3,058	28,167
X-linked	743	2/9	117	2,553	16/38	579
Total	42,043	269/881	6,129	128,642	1,604/3,096	28,746

In summary, we identified 49,322 coding variants in the 190 independent, European ancestry HSCR probands of which 8,506 were pathogenic (616 nonsense, 478 splice junction variants, 924 coding INDELS and 6,488 conserved (PhyloP $\geq$ 4) missense). Pathogenic variants were distributed across 5,271 genes.

Supplementary Table S4*: Exome sequence data accuracy.*

For quality control (QC) of these data we compared the sequences of six duplicate HSCR samples and assessed their concordance to those of 33 duplicate pairs in the 1000G data. The case samples were compared at between 777,945 and 896,652 sites with discordance varying between 7.4×10^{-5} and 3.24×10^{-4} in contrast to a discordance of $1.07 \times 10^{-4} \pm 4 \times 10^{-5}$ in controls. Therefore, the average discordancy rate is 1.57×10^{-4}.

Sample	***# sites***	***# missing***	***# concordant***	***# discordant***	***discordance rate***
HSCR2564	883,181	16,422	883,110	71	0.000080
HSCR18	777,945	121,658	777,794	151	0.000194
HSCR218	783,669	115,934	783,415	254	0.000324
HSCR1572	896,652	2,951	896,565	87	0.000097
HSCR430	892,302	7,301	892,236	66	0.000074
HSCR3685	890,554	9,049	890,403	151	0.000170

Table S5: *Sequence similarity between cases and their relatives.*

As a further check on data quality, we used the exome sequence data to assess the expected versus estimated genetic relationship between each affected and his/her sequenced relative. For these analyses, we used genotype data on 27,411 common autosomal exome variants for which allele frequencies were available from external controls (a subset of the data reported in **Supplementary Figure S1**). These estimates demonstrate the linear fit of observations to theoretical expectations (C. C. Li et al., 1993). From these results we estimated the coefficient of relationship as shown below (see **Supplementary Methods**):

Table S5.1: *Similarity measures using common variants.*

Expected coefficient of relationship (r)	***S***					
	Min.	***1st Quartile***	***Median***	***Mean***	***3rd Quartile***	***Max.***
0.5 (n = 41)	0.8910	0.9066	0.9079	0.9073	0.9092	0.9257
0.25 (n = 7)	0.8587	0.8627	0.8667	0.8661	0.8683	0.8750
0.125 (n = 4)	0.8338	0.8355	0.8367	0.8389	0.8400	0.8484
0.0625 (n = 1)	0.8248					
0 (n = 17, 955)	0.8025	0.8125	0.8146	0.8145	0.8166	0.8270

Table S5.2: *Coefficient of relationship corresponding to similarity in Table S5.1.*

Expected coefficient of relationship (r)	***R***					
	Min.	***1st Quartile***	***Median***	***Mean***	***3rd Quartile***	***Max.***
0.5 (n = 41)	0.4123	0.4962	0.5032	0.5002	0.5103	0.5995
0.25 (n = 7)	0.2381	0.2597	0.2810	0.2778	0.2899	0.3261
0.125 (n = 4)	0.1036	0.1132	0.1193	0.1312	0.1374	0.1827
0.0625 (n = 1)	0.0551					

Supplementary Table S6: *Pathogenic allele distribution in cases versus controls.*

Exome sequence analyses of the 190 HSCR cases identified 10 genes showing enrichment in the number of *distinct* SNVs, including *RET* and *EDNRB*, which serve as positive controls (**Table 2**). Based on gene expression studies in the human and mouse embryonic gut, all genes except *MMAA* were considered to be HSCR-relevant. The 7 novel genes identified had a total of 39 PAs (1 nonsense, 36 missense, 1 intronic and 1 exonic splicing change) which occurred in 40 of the 190 (21.1%) subjects. However, these cases also had additional PAs in the 17 previously identified HSCR genes (**Table S1**). For completeness, we list below by gene (column 1), the numbers of PAs (column 2) and the numbers of affected individuals with these PAs (column 3) in all 24 HSCR genes among the 190 cases (**Table S6**). The allele frequencies of these PAs as estimated from non-Finnish European ancestry subjects without a neuro-psychiatric disorder from ExAC (Exome Aggregation Consortium et al., 2016) are shown in column 4 (**Table S6**). These data from HSCR patients are compared to two types of controls. In the first, we compared the numbers of PAs (column 5) and the numbers of individuals (column 7) with these alleles, defined identically as in HSCR cases, among 740 non-Finnish European ancestry 1000G and NIMH controls (**Table S6**). In the second, we counted the numbers of PAs and cases only for alleles observed in cases (columns 6 and 8). Estimating these numbers from ExAC is not possible because we do not have access to the genotypes of individuals.

Cases in this study harbored 36 distinct PAs in 17 previously known, 39 PAs in 7 novel genes or a total of 75 distinct PAs in all 24 genes. These PAs occur in 41 (21.6%), 40 (21.1%) and 66 (34.7%) individuals for the known, novel and all HSCR genes. The mean allele frequencies of these PAs in our sample of HSCR cases for known, novel and all HSCR genes are 5.58×10^{-4}, 2.96×10^{-4} and 4.22×10^{-4}, respectively, showing relatively little difference between these three categories, but being ~12 times larger than identically defined PAs reported for known genes in databases (i.e., 3.61×10^{-5}) (**Table S1**). We suspect that this is owing to the selective reporting of more severe and rarer alleles in public databases and missing true disease variants of lower penetrance which are expected to have higher allele frequencies.

(i) We first tested whether our definition of PAs enriches for causal alleles among HSCR cases in the identified genes as compared to the 740 1000G and NIMH controls. In controls, we identified 29 distinct PAs in the 17 previously known genes, 23 PAs in 7 novel genes and 52 PAs in all 24 genes, and these occurred in 71 (9.6%), 28 (3.8%) and 98 (13.2%) controls, respectively. Overall, identically defined PAs were identified in 34.7% (66/190) of cases in contrast to 13.2% (98/740) of controls, demonstrating a 2.63-fold enrichment (2-sided: $P = 4.08 \times 10^{-12}$). This enrichment was evident for both known genes (41/190 in cases versus 71/740 in controls: 2.25-fold, $P = 5.97 \times 10^{-6}$) and, necessarily (identified using this criterion), novel genes (40/190 in cases versus 28/740 in controls: 5.56-fold, $P = 3.46 \times 10^{-16}$). Observe also that in controls, these PAs had average ExAC allele frequencies of 1.11×10^{-3}, 4.74×10^{-4}, 8.26×10^{-4}, in known, novel and all HSCR genes respectively, and were ~2-fold *higher* than the corresponding allele

frequencies in HSCR patients (2-sided: $P = 2.14 \times 10^{-5}$, 0.066 and 1.62×10^{-5}, respectively, for known, novel and all HSCR genes). Thus, our definition of PAs leads to enrichment of causal variants because these selected variants exist in significantly greater numbers in cases than in controls and they are significantly rarer in the population than similarly defined variant alleles in controls. Note that 98 of 740 controls or 13.2% of controls have PAs: these represent both non-causal alleles and low penetrance disease alleles unobserved in our cases. Thus, we have a maximum false positive rate of 13.2% in identification of causal alleles in cases. The true proportion of falsely identified PAs in cases is, however, much lower because causal alleles are enriched in cases. In any case, we have significant statistical evidence of an enrichment of HSCR causal alleles across all 24 genes.

(ii) Given the effects of the 24 genes in HSCR, we assessed the impact of observed variants at these genes by performing direct association tests of variant frequencies in cases and controls by gene, i.e., we restricted attention to only PAs observed in cases. We observed 12 case-specific PAs (6 each for known and novel HSCR genes) among 37 (5.0%) of 740 controls (29 and 8 individuals for known and novel HSCR genes). Across all 24 genes, this number is significantly smaller than the 75 among HSCR cases ($P = 9.15 \times 10^{-55}$) and they occur in 37 controls which is also considerably smaller than that in the 66 HSCR cases ($P = 2.27 \times 10^{-31}$). The number of PAs identified in HSCR cases is 24-fold higher than in controls, and the number of individuals with such alleles is 7-fold greater than in controls. These significant differences are true for both known and novel genes as a group. We do not have statistical power to assess these effects for individual genes but the results can be accumulated by pathways (see **Table S1**), as in the following **Table S7**, so that the *relative contributions* of different gene classes to HSCR risk can be estimated.

Gene	190 cases			740 controls				
	# unique PAs	# cases with PAs	average ExAC allele frequency by gene	# unique PAs		# controls with PAs		average ExAC allele frequency
SOX10	1	1	0	0[a]	0[b]	0[a]	0[b]	-
PHOX2B	1	1	0	0	0	0	0	-
ZEB2	2	2	0	2	0	2	0	3.57×10^{-5}
TCF4	0	0	-	0	0	0	0	-
GDNF	0	0	-	0	0	0	0	-
NRTN	0	0	-	0	0	0	0	-
GFRA1	1	1	7.16×10^{-5}	1	1	1	0	2.35×10^{-5}
RET	9	12	3.33×10^{-4}	3	0	5	3	1.18×10^{-3}
ECE1	0	0	-	2	0	2	0	9.79×10^{-5}
EDN3	1	1	0	1	0	4	0	2.10×10^{-3}
EDNRB	7	7	0	1	0	1	0	2.37×10^{-5}
KIF1BP	1	1	7.18×10^{-5}	3	0	5	0	1.40×10^{-4}
L1CAM	1	1	0	0	0	0	0	-
IKBKAP	1	1	1.00×10^{-3}	5	1	6	1	2.69×10^{-4}
SEMA3C	3	3	2.10×10^{-3}	3	1	25	10	4.28×10^{-3}
SEMA3D	6	8	9.25×10^{-4}	4	2	9	7	1.83×10^{-3}
NRG1	2	3	2.05×10^{-3}	4	1	11	8	1.04×10^{-3}
ADAMTS17	5	5	2.95×10^{-5}	1	0	1	0	4.00×10^{-4}
ACSS2	6	6	2.37×10^{-4}	2	1	2	1	6.00×10^{-4}
SLC27A4	4	4	1.55×10^{-4}	1	0	1	0	2.00×10^{-4}
SH3PXD2A	4	4	4.62×10^{-4}	2	1	4	1	2.50×10^{-4}
ENO3	5	5	1.50×10^{-4}	2	0	2	0	3.60×10^{-5}
FAM213A	4	6	7.30×10^{-4}	1	1	2	2	2.80×10^{-3}
UBR4	11	15	3.50×10^{-4}	14	3	17	4	4.10×10^{-4}
All Genes	**75**	**66**	**4.22×10^{-4}**	**52**	**12**	**98**	**37**	**8.26×10^{-4}**

[a]*: number of distinct PAs in controls;* [b]*: same as in* [a] *but restricted to alleles observed in cases*

Table S7: *Distribution and effect of case-observed PAs by pathway.*
These are data in **Table S6** rearranged by gene class (defined in **Table S1**) with statistically significant odds ratios in bold.

Pathway	***Genes***	***# cases (n = 190)***		***# controls (n = 740)***		**Pathway odds ratio (95% CI)**
RET	*GDNF*	0	13	0	3	**16.03** (5.21-49.28)
	NRTN	0		0		
	GFRA1	1		0		
	RET	12		3		
EDNRB	*ECE1*	0	8	0	0	**68.98** (8.68-547.92)
	EDN3	1		0		
	EDNRB	7		0		
SEMA3	*SEMA3C*	3	11	10	17	**2.65** (1.25-5.60)
	SEMA3D	8		7		
TFs	*SOX10*	1	4	0	0	**35.73** (4.15-307.72)
	ZEB2	2		0		
	PHOX2B	1		0		
	TCF4	0		0		
remaining genes	*KIF1BP*	1	7	0	9	**3.15** (1.22-8.09)
	L1CAM	2		0		
	IKBKAP	1		1		
	NRG1	3		8		
17 known genes	*all the above*	*41*		*29*		**6.70** (4.06 – 11.04)
novel genes	*ADAMTS17*	5	40	0	8	**23.19** (11.04-48.72)
	ACSS2	6		1		
	SLC27A4	4		0		
	SH3PXD2A	4		1		
	ENO3	5		0		
	FAM213A	6		2		
	UBR4	15		4		
All 24 genes	*all the above*	*66*		*37*		**10.02** (6.45 – 15.58)

[1] Odds ratios were calculated using the Haldane bias correction and by comparing 190 cases with 740 controls based on coding PAs observed in cases only.

Table S8: *Identifying CNVs using exome sequence, SNP array and karyotype data.*

Details of each CNV detected and validated, based on multiple data types, are shown with CNV location, type, size, chromosomal locus and observed numbers in 185 cases and 19,584 controls. We separately validated a 4 kb *RET* deletion (chr10:42917793-42922026) in patient HSCR472 from a low-quality CNV.

Sample ID	Karyotype, CNV		CNV size (kb)	CNV Location[1]	Karyotype, Microarray results	SNP Validation[2]	Observed #		P value[4]
	State	Chr.					Case	Control[3]	
many	dup	21	47,710	1-46,709,983	47 XX & XY, +21	+	11	17[4]	$6.68 \times 10^{-16*}$
HSCR2970	del	16	985	28299106-29283882	16p11.2 del	+	3	12	$3.38 \times 10^{-4*}$
HSCR4220	del	16	906	29372452-30278662	-	+			
HSCR71	del	16	740	29372452-30112616	-	+			
HSCR4522	dup	1	509	144126136-144634799	-	+	3	27	2.72×10^{-3}
HSCR4584	dup	1	1,185	143565872-144750520	-	+			
HSCR46	dup	1	971	143663355-144634799	-	+			
HSCR3886	del	1	1,425	144751127-145931774	-	+	1	6	$6.37 \times 10^{-2*}$
HSCR491	del	22	8,000	16280000-24230000	22q11.2 del	+	1	0	$9.36 \times 10^{-3*}$
HSCR3186	dup	22	1,447	15826987-17273998	Tetrasomy 22q	+	1	0	$9.36 \times 10^{-3*}$
HSCR522	dup	17	1,835	13340236-15175628	-	+	1	5	$5.49 \times 10^{-2*}$
HSCR11	dup	4	7,768	183482167-191247414	47, XX, +der(15)t(4:15)	+	1	0[5]	$9.36 \times 10^{-3*}$
HSCR11	dup	15	3,800	61487053-65287054					
HSCR73	del	1	582	46916717-47498799	-	+	1	0	9.36×10^{-3}
HSCR208	del	12	554	8102597-8656675	-	+	1	0	9.36×10^{-3}
HSCR4368	del	13	14,356	70912755-85268645	13q21.33-q31.1 del	+	1	0	$9.36 \times 10^{-3*}$
HSCR3305	del	2	8,847	133437762-142284441	-	+	1	0	$9.36 \times 10^{-3*}$
HSCR423	del	8	579	1501255-2080313	-	+	1	0	9.36×10^{-3}
HSCR241	dup	2	1,377	31566-1397283	-	+	1	1	1.86×10^{-2}
HSCR500	dup	7	1,498	88227224-89725429	-	+	1	11	1.06×10^{-1}
HSCR4178	inv	10	25,600	101900000-127500000	10q24.3-q26.13inv	-	1	0[5]	-

[1]hg18 genome coordinates; [2] Human Omni 2.5-4 v1 BeadChip data; [3]observed numbers (50% reciprocal overlap of each CNV) in 19,584 controls from (Coe et al., 2014); [4] controls used were ascertained as adults and not expected to include trisomy 21, the rate of which in 19,584 births was estimated from population studies (Presson et al., 2013). [5]Note that the array studies in controls could not detect aneuploidies, translocations and inversions. The control counts for 47, XX, +der(15) t(4:15) are for the two duplications at the translocation site; for the 10q24.3-q26.13 inversion, control counts were not available and not expected. P-values with an asterisk indicate pathogenic CNVs as designated in Table S9.

Table S9: *Inferring the phenotypic consequences of karyotype variants and CNVs.*

Karyotype/ copy number variant[1]	***Syndrome***	***P[2]***	***Assessment of Causality[3,4]***
Free & mosaic trisomy 21	Y	**6.68×10^{-16}**	Pathogenic – known association (HSCR)
16p11.2 del	Y/2N	**3.38×10^{-4}**	Pathogenic – known association (DD)
1q21.1 dup	N	**2.72×10^{-3}**	Likely benign
1q21.1 del	Y	6.37×10^{-2}	Pathogenic – known association (DD)
22q11.2 del	Y	9.36×10^{-3}	Pathogenic – known association (DD)
Tetrasomy 22q	Y	9.36×10^{-3}	Pathogenic – known association (cat eye)
17p11.2 dup	N	5.49×10^{-2}	Pathogenic - known association (CMT1A)
47, XX, +der(15) t(4:15)	Y	9.36×10^{-3}	Pathogenic – large duplication with 4q partial trisomy
1p33 del	N	9.36×10^{-3}	VOUS
12p13.31 del	Y	9.36×10^{-3}	VOUS (large segmental duplication content)
13q21.33-q31.1 del	Y	9.36×10^{-3}	Pathogenic – known association (DD)
2q21.2-q22.2 del	Y	9.36×10^{-3}	Pathogenic – known association (DD)
8p23.3 del	Y	9.36×10^{-3}	VOUS (genes in interval have deletions in controls)
2p25.3 dup	N	1.86×10^{-2}	Likely benign
7q21.12 dup	N	1.06×10^{-1}	Benign
10q24.3-q26.13 inv	Y	-	VOUS

[1] CNVs of interest were defined as deletions >500kb and duplications >1 mb with a control frequency of <1% (Kaminsky et al., 2011). Considering all of these CNVs of interest (listed in Table S8) except for the 10q24.3-q26.13 inversion (because a control frequency could not be determined), we observed a total of 29 cases (of 185) having a CNV of interest compared to an expected control frequency of 700 (of 19,584) corresponding to an odds ratio of 5.10 (95% CI: 3.43 – 7.57, $P = 4.27 \times 10^{-11}$). However, most of these changes in controls do not have a phenotypic effect and were assessed against primarily known causal changes, which is why we decided to use only a smaller set of known pathogenic variants for risk estimation. [2]Entries in bold are statistically significant after multiple (15) test corrections with overall significance level of 5%. [3] The presence of a CNV in a HSCR patient can be a causal event or an incidental finding. We assessed known CNV-HSCR associations, statistical evidence of a new CNV association (column 3) and previous CNV association with a developmental phenotype from a set of 29,085 cases of developmental disorders (DD) (Coe et al., 2014; Kaminsky et al., 2011), for assessing CNV pathogenicity. [4] VOUS is variant of unknown significance. We ultimately classified variants as "pathogenic" based on a known association with a developmental disorder; these pathogenic CNVs include Free and mosaic trisomy 21, 16p11.2 del, 1q21.1 del, 22q11.2 del, tetrasomy 22q, 17p11.2 dup, 47, XX, +der(15) t(4:15), 13q21.33-q31.1 del and 2q21.2-q22.2 del.

Table S10: *Comparison of genetic burden of classes of variation by sex.*

Disease-associated risk allele class		% frequency		Male odds ratio (95% CI)	Female odds ratio (95% CI)
		cases (M/F)	controls		
Enhancers, common variants[1]	known[4]	54.2/38.2	17.1	5.78 (4.01-8.33)	3.05 (1.86-5.00)
Coding genes, rare variants[2]	known & novel[4]	35.2/27.9	5.0	10.32 (6.41-16.63)	7.46 (4.09-13.60)
	known[4]	23.0/16.2	3.9	7.31 (4.22-12.65)	4.86 (2.36-10.04)
Copy number alterations, rare variants[3]	known & novel[4]	14.2/6.2	0.20	81.99 (45.51-147.70)	35.60 (13.05-97.13)
	known[4]	8.3/1.5	0.09	100.95 (46.36-219.83)	24.79 (4.61-133.21)

[1] Five or more common disease variants (Table 1) were observed in 90 of 186 cases and 107 of 627 controls; [2] rare coding sequence variants (Table 2) were identified in 66 of 190 cases with an expected rate of 37 in 740 controls; [3] copy number variants (Table 3) were identified in 21 of 185 cases with an expected rate of 40 in 19,584 controls. [4] The data relevant to 24 known and novel loci, and the 18 known loci, respectively, are shown subdivided by sex and are the same data as in Table 4 of the main paper.

Table S11. *Distribution of HSCR by mutation type and phenotype.*

Common Variant[a]	***Rare Variant[b]***	***CNV[c]***	***# (%) Cases[d]***	***# (%) Male / Female***	***# (%) Short / Long & TCA[e]***	***# (%) Simplex / Multiplex***	***# (%) non-syndromic / MA[f]***
-	-	-	50 (28)	26 (52) / 24 (48)	17 (46) / 20 (54)	28 (56) / 22 (44)	38 (76) / 12 (24)
+	-	-	53 (30)	35 (66) / 18 (34)	24 (57) / 18 (43)	36 (68) / 17 (32)	46 (87) / 7 (13)
-	+	-	27 (15)	14 (52) / 13 (48)	9 (41) / 13 (59)	14 (52) / 13 (48)	17 (63) / 10 (37)
-	-	+	13 (7)	11 (85) / 2 (15)	9 (82) / 2 (18)	12 (92) / 1 (8)	2 (15) / 11 (85)
+	+	-	29 (16)	24 (83) / 5 (17)	14 (58) / 10 (42)	24 (83) / 5 (17)	20 (69) / 9 (31)
+	-	+	1 (1)	1 (100) / 0 (0)	0 (0) / 1 (100)	1 (100) / 0 (0)	1 (100) / 0 (0)
-	+	+	3 (2)	3 (100) / 0 (0)	2 (67) / 1 (33)	2 (67) / 1 (33)	0 (0) / 3 (100)
+	+	+	3 (2)	1 (33) / 2 (67)	2 (67) / 1 (33)	3 (100) / 0 (0)	0 (0) / 3 (100)
Totals			179 (100)	115 (64) / 64 (36)	77 (54) / 66 (46)	120 (67) / 59 (33)	124 (69) / 55 (31)

[a] Common variant: 5 or more risk alleles at *RET* (rs2435357, rs2506030, rs7069590) and *SEMA3D* (rs11766001); [b] Rare Variant: 1 or more rare, deleterious variants in any of 17 known and 7 new susceptibility genes identified in this study; [c] CNV (copy number variant): a clinically identified alteration (trisomy 21, 22q deletion, etc.), recurrent CNV or unique rare deletion >500kb or duplication >1000kb identified as pathogenic in Table S9; [d] 179 affected individuals with complete data for all three mutation classes; [e] Cases where segment length was uncertain have been excluded here; [f] Non-syndromic cases have no clinical diagnosis of recognized syndromes or multiple anomalies (MA) in addition to HSCR.

Table S12: *Functions of novel HSCR genes and their relevance to ENS development.*

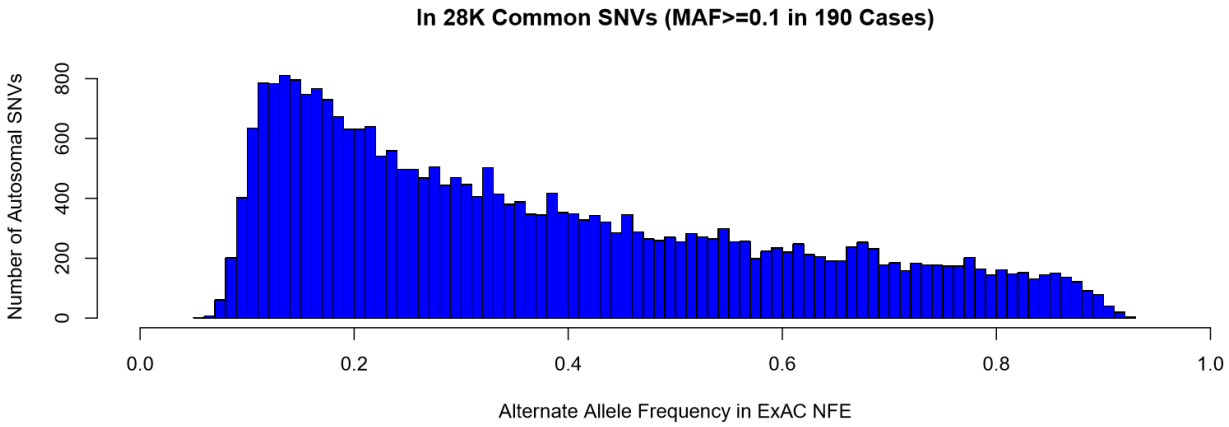
Relevance to ENS development.	***Gene***	***Encoded functions***
Regulation of axonal guidance	*ADAMTS17*	A plasma membrane protein whose knockdown induces breast cancer cell apoptosis; acts as a versicanase in development and is dysregulated by epigenetic alterations (Jiang et al., 2015; Kelwick et al., 2015).
	SH3PXD2A	A lipid-binding cytoskeletal protein resident in the embryonic mesenchyme, binds many ADAM proteins and functions to locally degrade extracellular matrix during axon guidance through tissues. Analysis of zebrafish embryos and neural crest cells *in vitro* have indicated that Src-activated Tks5 (protein encoded by *SH3PXD2A*) is necessary for proper neural crest cell migration (Murphy et al., 2011).
Cell growth & proliferation	*ACSS2*	Acetyl-Coenzyme A synthetase 2 is both cytoplasmic and nuclear. Despite having many functions in lipid synthesis and energy generation, it can affect transcriptional control and gene expression through p300-catalyzed control of histone acetylation versus crotonylation (Sabari et al., 2015).
	SLC27A4	A fatty acid transport protein localized to the endoplasmic reticulum and the plasma membrane which has acyl-CoA ligase activity and, therefore, could have functions that interact with *ACSS2*, since increased fatty acid synthesis is required to meet the demand for membrane expansion of rapidly growing cells.
	UBR4	A ubiquitin E3 protein ligase (component N-Recognin 4) localized to the cytoskeleton and the nucleus. Despite having a function required for the termination of RET signaling (performed by CBL (Mulligan, 2014)), UBR4 may also be involved in regulating acetylation versus ubiquitylation by competing for the same lysine residues in the regulation of fatty acid synthesis and cell growth (Lin et al., 2013).
	ENO3	Encodes a muscle-specific enolase active during development.
Local inflammation	*FAM213A*	A cytoplasmic and mitochondrial redox-regulatory protein. Recently, sulfhydryl-mediated redox signaling in inflammation has been shown to have a significant role in neuro-degenerative diseases using RET target proteins (Gorelenkova Miller & Mieyal, 2015).

Table S13: Translation blocking morpholinos for zebrafish orthologs of HSCR associated genes

Gene	***Transcript id***	***Morpholino sequence***
Control	-	CCTCTTACCTCAGTTACAATTTATA
Ret	NM_181662	ACACGATTCCCCGCGTACTTCCCAT
ednrba	NM_131197	GGAAACGCATGACTATTTAACAGTC
ednrbb	XM_683473.5	GCAGCAGAATGACCGATGATGCCAT
ubr4	XM_005162190	CTCCATCTCCTCCACTCGACGCCAT
eno3	NM_214723	GCGTGAATCTTACTAATGGACATCC
mmaa	NM_001105112	AAAACTCTAGATGGACGCATCTTTC
sh3pxd2aa	NM_001160022	TTGGGAACTTGTCGAGTATCTGCAT
slc27a4	NM_001017737	TGGCACACGCCAACCGCAACATCCT
acss2	NM_001002641	CAATCAGAGAGTGCCAACACATATC
fam213aa	NM_001193525	CAAGGCCAAGTGACCACATGCCCAT

3.7.3 Supplementary Figures

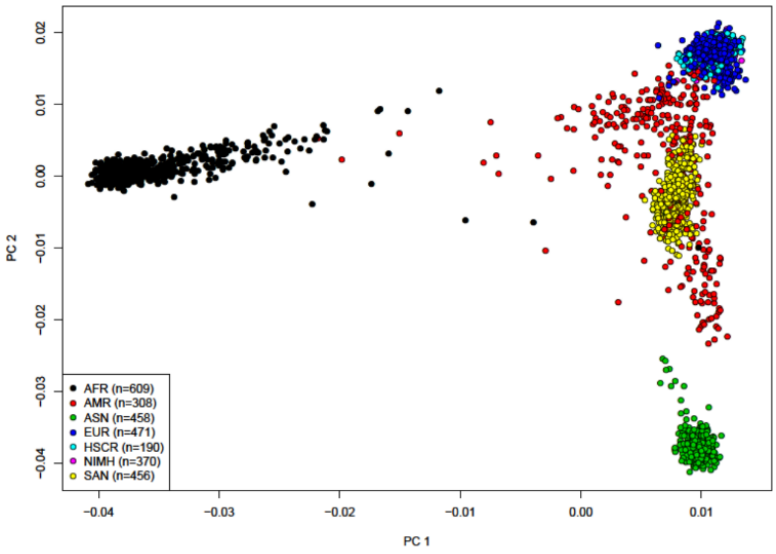
Supplementary Figure S1: *Allele frequency distribution of 28,746 common autosomal variants among the 190 HSCR cases* (see Table S3).



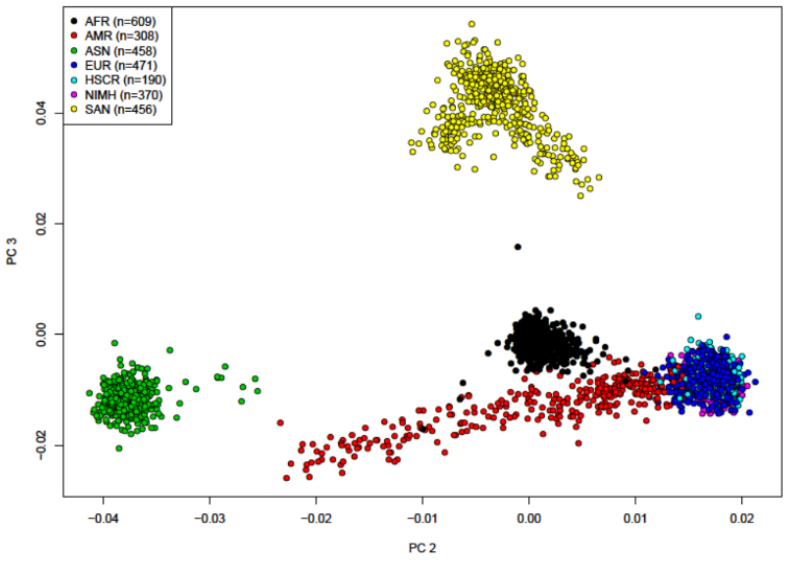
Supplementary Figure S2: *Principal component analysis (PCA) of HSCR samples.*

In A and B, the first three PCs are plotted for PCA of 190 HSCR non-Mennonite independent cases (HSCR NI); 370 European American samples from NIMH (NIMH); 458 East Asian samples from 1000G (ASN); 471 European samples from 1000G (EUR); 609 African samples from 1000G (AFR); 308 American samples from 1000G (AMR); 456 South Asian samples from 1000G (SAN). The results show clear overlap for all 190 HSCR cases with reference individuals of European ancestry. PCA of Europeans only (first three PCs plotted in C and D) showed that the HSCR cases cannot be distinguished from any European ancestry group except the Finns.

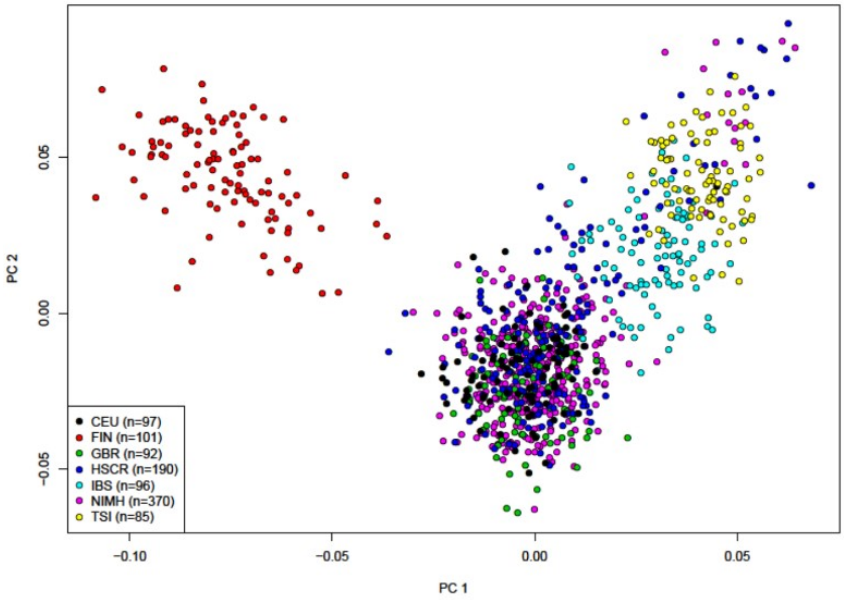
A



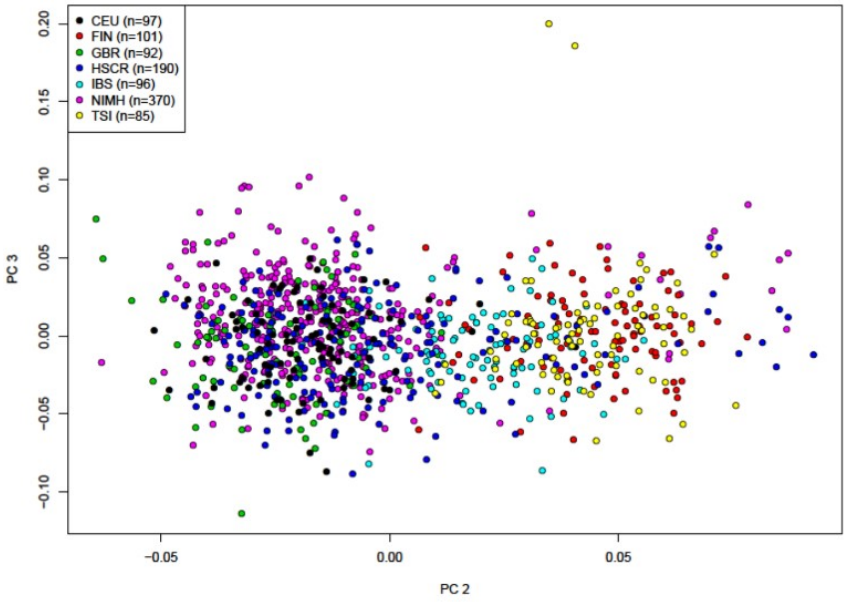
B



C

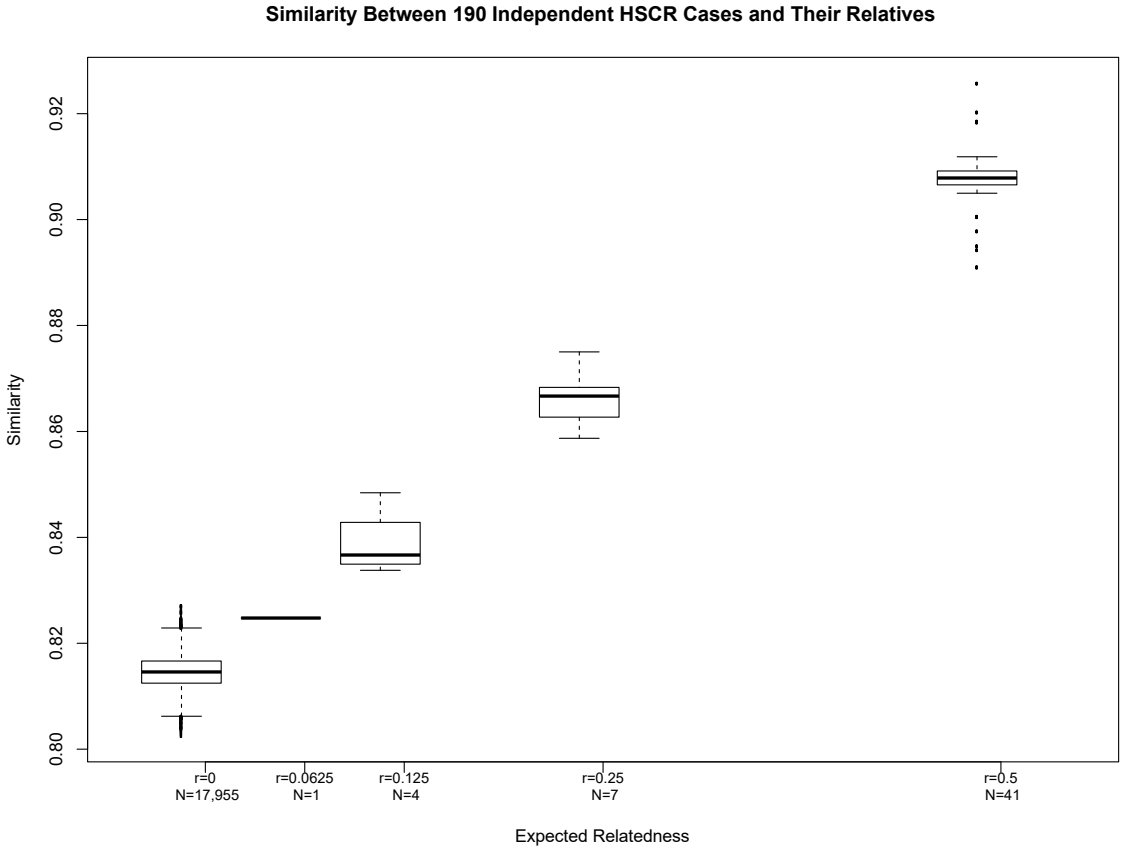

CEU (n=97)
FIN (n=101)
GBR (n=92)
HSCR (n=190)
IBS (n=96)
NIMH (n=370)
TSI (n=85)
PC 1
PC 2
−0.10
−0.05
0.00
0.05
−0.05
0.00
0.05

D


CEU (n=97)
FIN (n=101)
GBR (n=92)
HSCR (n=190)
IBS (n=96)
NIMH (n=370)
TSI (n=85)
PC 2
PC 3
−0.05
0.00
0.05
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

Supplementary Figure S3: *Sequence similarity between relatives.*

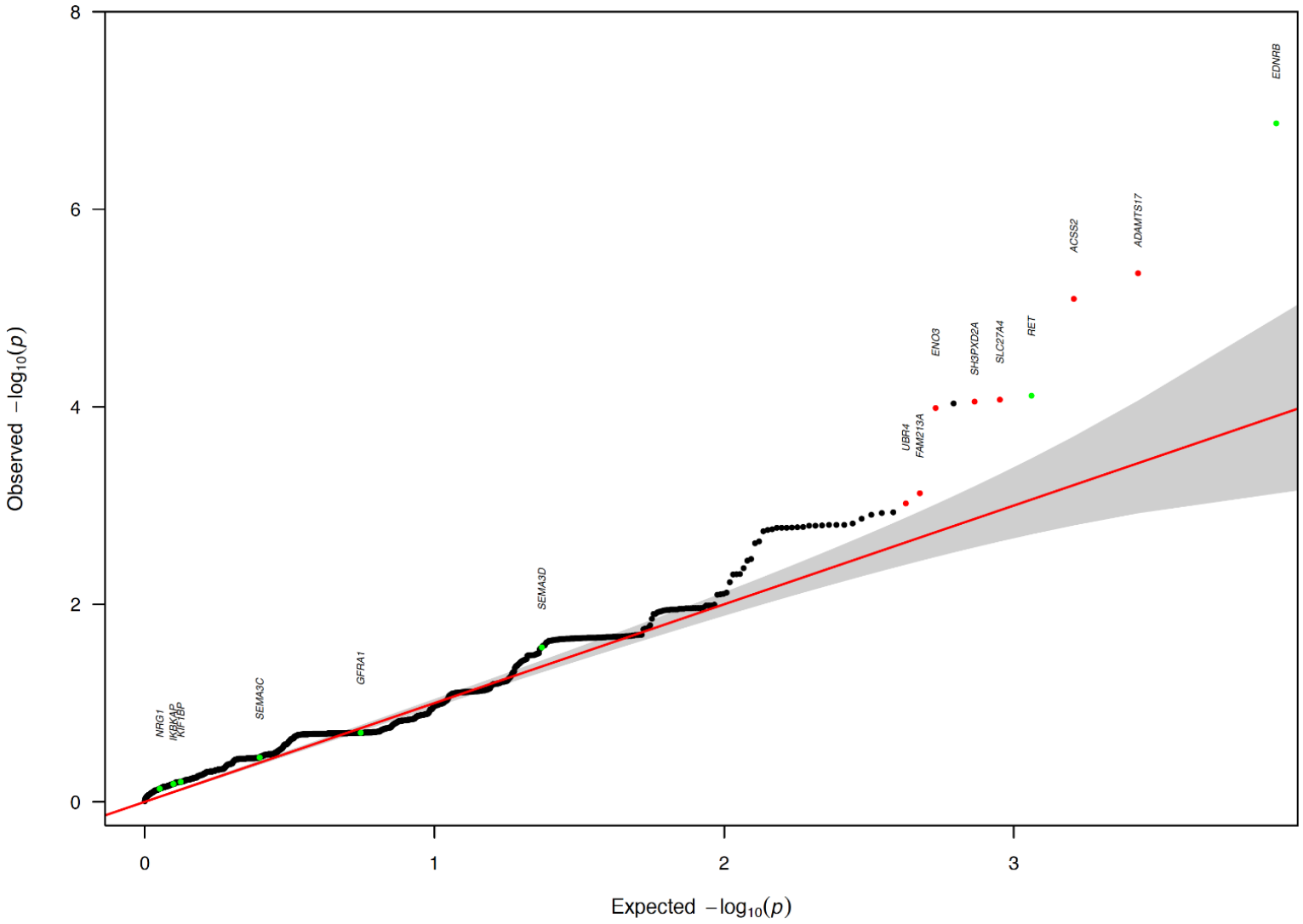
The distribution of similarity scores (S) for the expected (pedigree-based) degree of relationship is summarized below (see data in **Tables S5**). S is linearly related to the coefficient of relationship, as expected, verifying the putative relationships with genetic data.



Supplementary Figure S4: *Assessment of genes significantly enriched for PAs.*

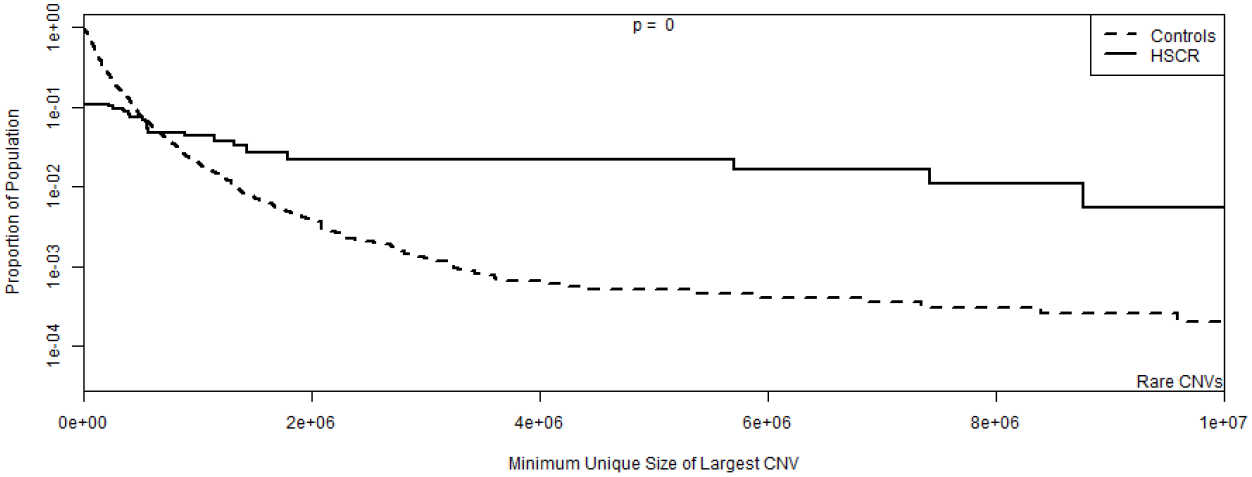
We used computer simulations, using the control exome sequence data, to compare the observed to expected distribution of distinct pathogenic alleles (PA) for each of 4,027 genes with at least one such variant in cases and controls. These were compared to their observed numbers in cases and are compared in the QQ plot below with a 95% confidence interval at each point. As explained in the main text, the top 10 genes were enriched as a group (P<0.001). Genes marked in green were previously identified HSCR genes and those marked in red are novel genes identified in this study.

The statistical test for comparing observed to expected numbers of distinct PAs assumed a Poisson distribution of the number of distinct PAs in a sample. This is a conservative assumption because comparisons of the variance to the mean of the number of distinct PAs in 190 samples, as assessed from replicate sampling from controls, shows considerably less-dispersion (see **Supplementary Figure S10**). The same statistical method was used to identify candidate HSCR genes from small INDELs. The test was applied to rare (MAF $\leq$ 0.05 in 190 cases or 740 controls) and common (MAF>0.05 in cases or controls) alleles for small insertions and deletions separately. There were 551 genes with rare small INDELs in both cases and controls but only one gene, *FAN1*, had a P value below 0.01. None of the 132 genes with common small INDELs showed any statistical significance. This is unsurprising given that most genes have very few (at most 3 rare and 2 common) INDELs.


Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
6
8
0
1
2
3
NRG1
IKBKB
KIF1BP
SEMA3C
GFRA1
SEMA3D
UBR4
FAM213A
ENO3
SH3PXD2A
SLC27A4
RET
ACSS2
ADAMTS17
EDNRB

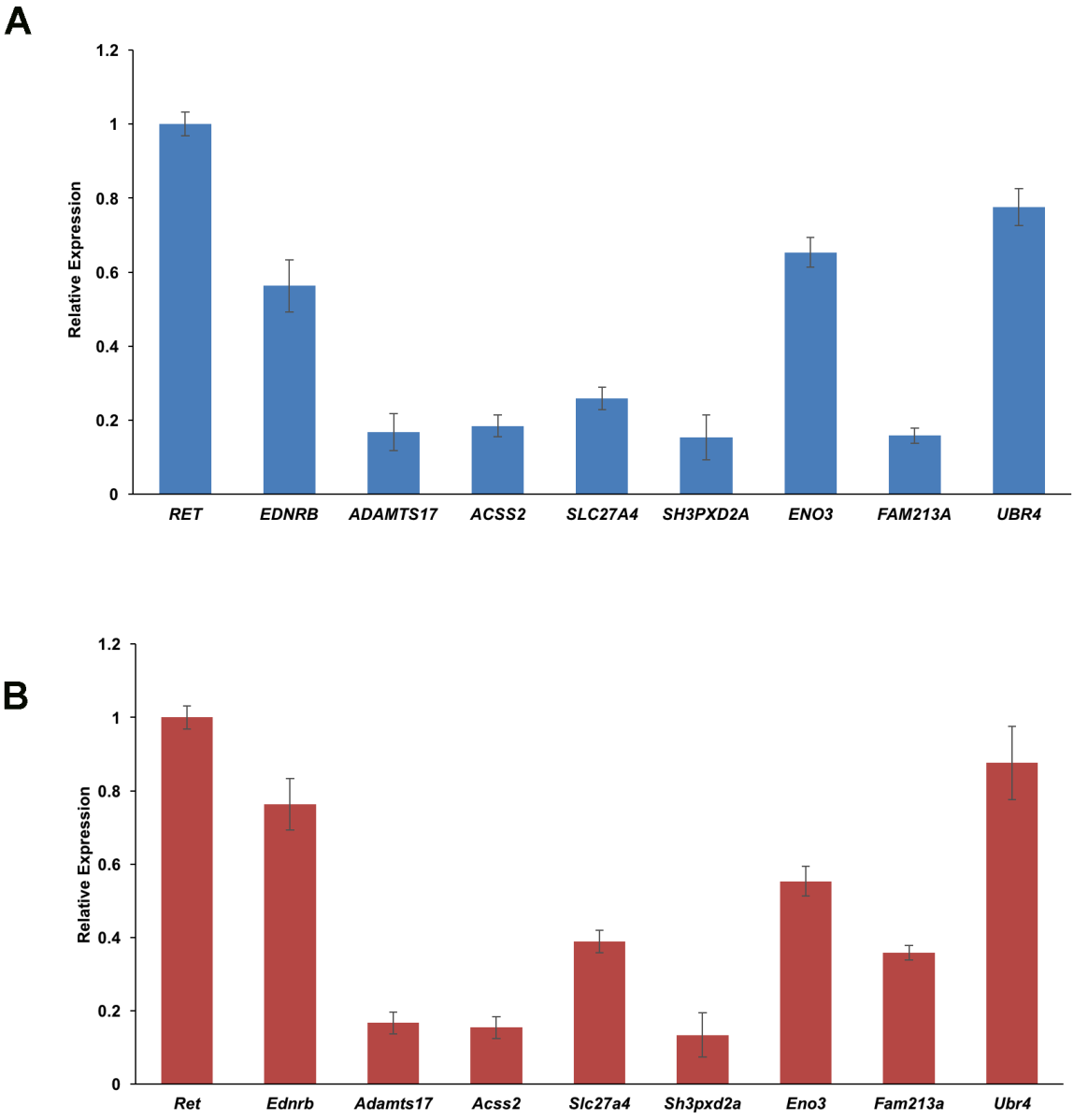
Supplementary Figure S5: *CNV burden in HSCR.*

The proportion of samples with any CNV, in either HSCR or controls, is plotted against the minimum unique size of the largest CNV. The data shows that the distribution of CNVs in HSCR is significantly greater ($P<2.2x10^{-16}$) than in controls by both the log-rank test and the Peto and Peto modification of the Gehan-Wilcoxon tests (https://stat.ethz.ch/R-manual/R-devel/library/survival/html/survdiff.html) (Harrington & Fleming, 1982). The lines cross at 500 kb. Note that CNV size in this analysis is corrected for segmental duplication content.



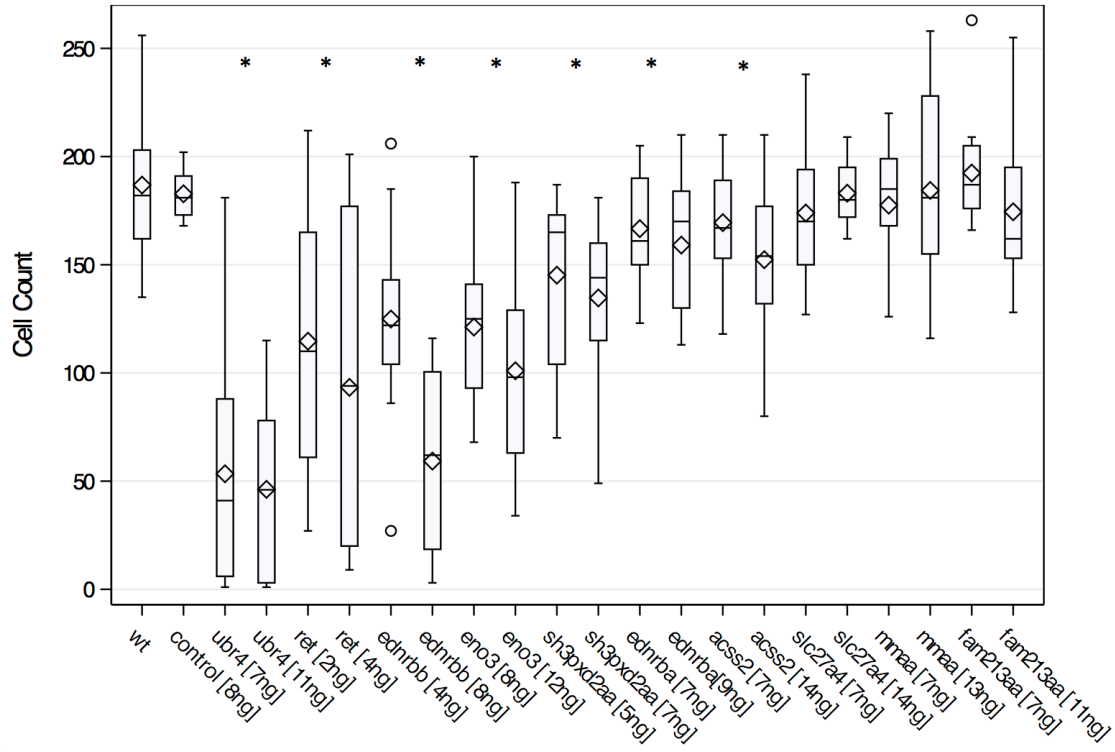
Supplementary Figure S6: *Gene expression of candidate HSCR genes in the embryonic human and mouse gut.*

Taqman gene expression profiles in human fetal gut tissue at Carnegie stage 22 shows all genes except *MMAA* are expressed at the relevant time point in development (A), with similar data from mouse gut tissues at E10.5 (B). The transcript with the highest expression was set to unity to compare the relative expression of other genes. The error bars represent standard errors of the mean from multiple measurements.

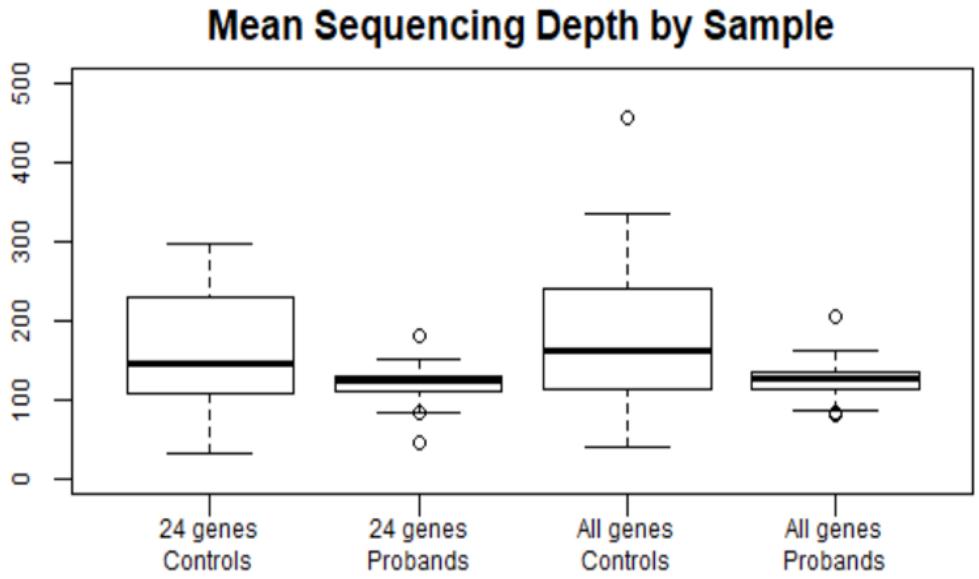


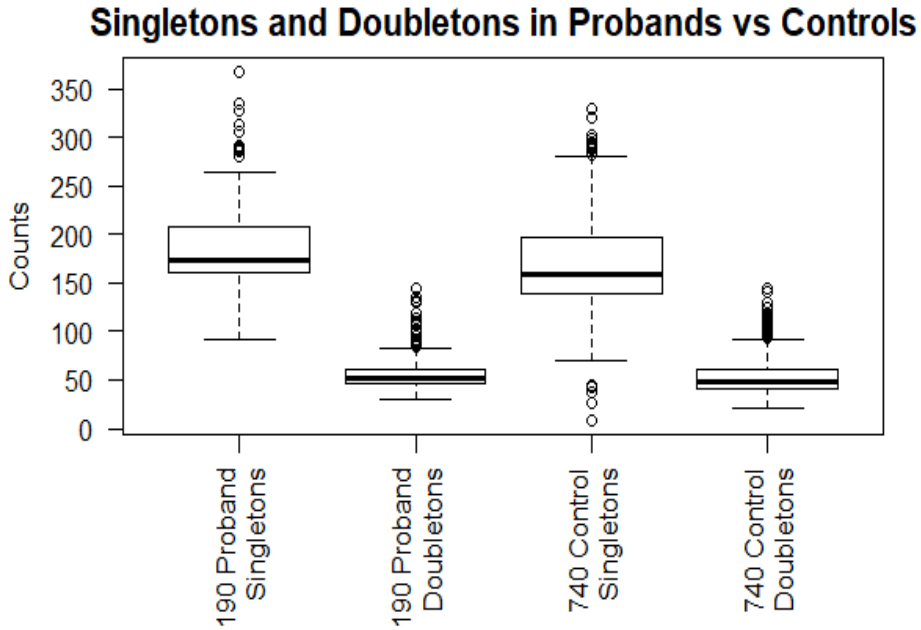
Supplementary Figure S7: *Assessment of HSCR candidate genes in zebrafish*.

Distribution of HuC positive migratory enteric neuronal precursors in 6 dpf zebrafish embryos from controls and knockdown of HSCR candidate gene orthologs. Genes with a statistically significant reduction in cell numbers are indicated by an asterisk. Note that there are two ednrb zebrafish orthologs but only ednrbb was significant in these assays; further, acss2 was significant only at the higher concentration.

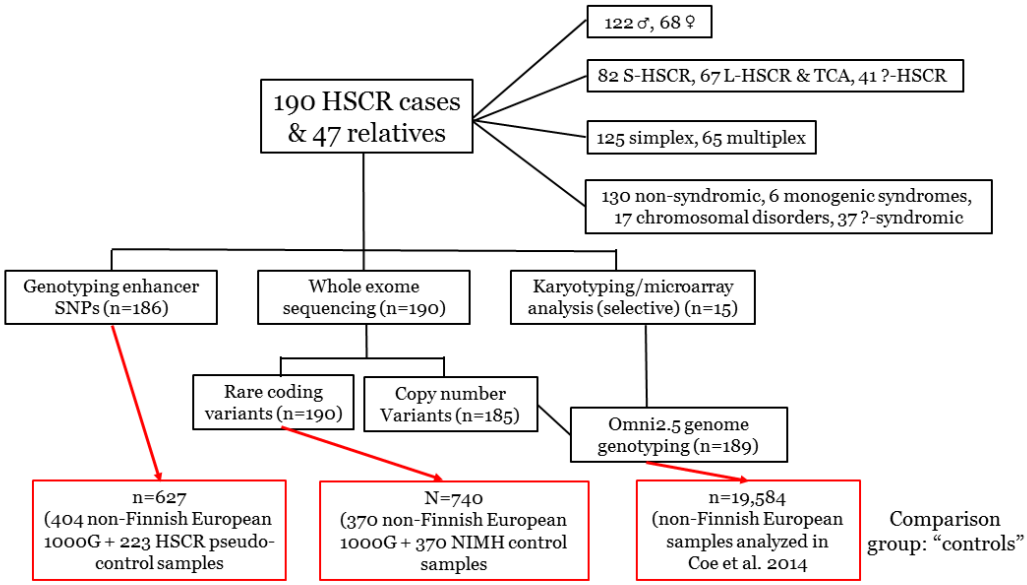


Supplementary Figure S8: *Comparison of depth of sequencing at HSCR genes and rare variant counts per individual in exome sequenced cases and controls*

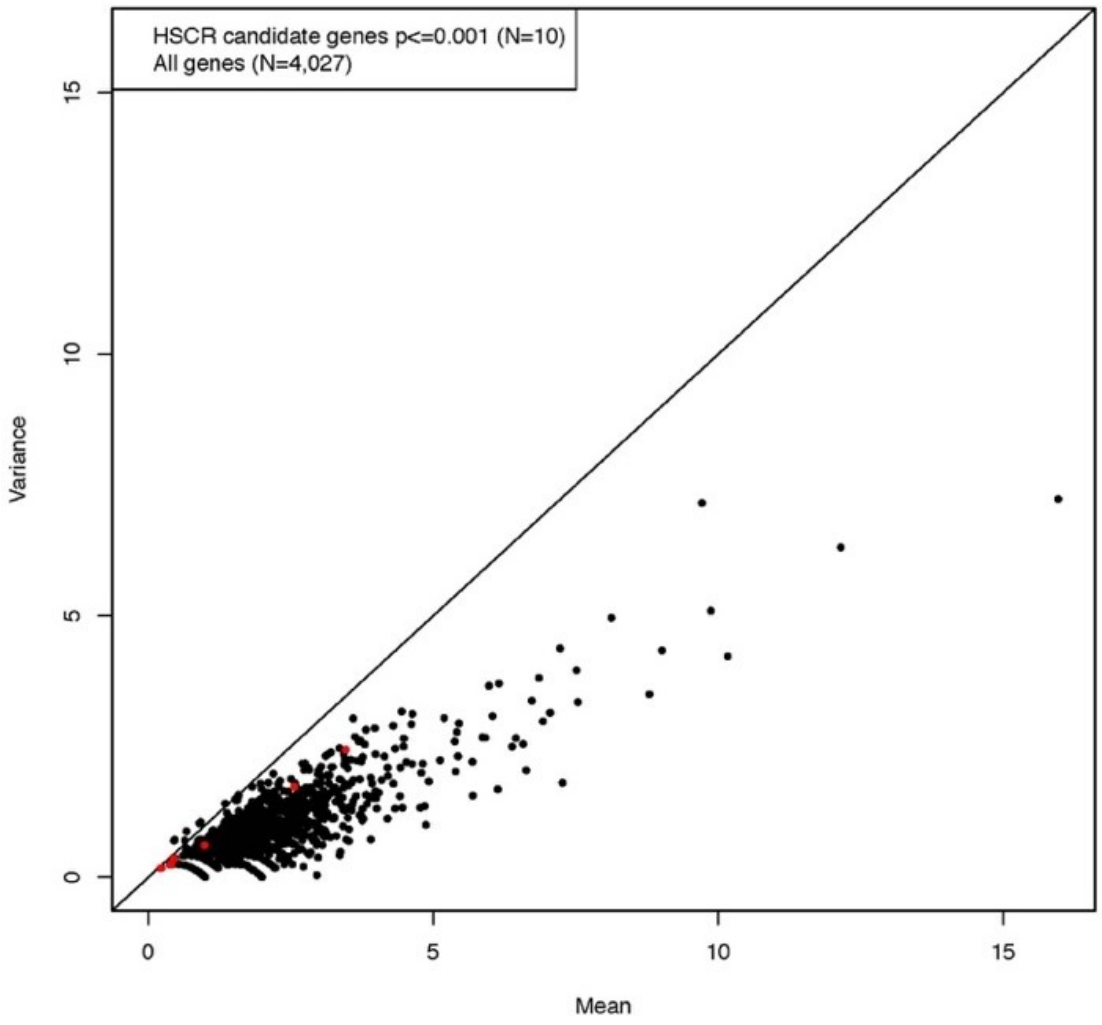




Supplementary Figure S9: *Overview of all case and comparison group samples analyzed, their sample sizes and the types of genetic analyses conducted on each*



Supplementary Figure S10: *Variance vs. mean of deleterious SNV counts per gene in 190 controls over 10,000 sampling events; N=4,027 genes with >= 1 del SNVs in both 190 cases and 740 controls*



Chapter 4: The genetic risk profile of female enriched multiplex families (FEMFs)

4.1 Introduction

It is well established that autism spectrum disorder (ASD) has a male skewed sex ratio of ~4:1 (Fombonne, 2003; Loomes et al., 2017). This higher disease prevalence in males can be explained by a multiple threshold liability model in which females are protected by not yet understood sex-biased mechanisms that raise their liability threshold for clinical disease. Under this model, females are affected only when they harbor a greater number of and or higher severity autism risk variants than their male counterparts. Consistent with this model, dizygotic twins of female probands show higher autism concordance than those of male probands (Robinson et al., 2013), and children with one affected older sibling have a higher recurrence risk when the older sibling is female than male – 7% for female and 3.7% for male older siblings (Jorde et al., 1991). Werling and Geschwind (2015) showed that the difference in recurrence risk is more pronounced between siblings having at least two affected older siblings – 44.3% for female-containing and 30.4% for male-only proband pairs. They further showed that a shorter inter-birth interval increases recurrence risk specifically in male-only families, indicating greater environmental risk for male-only families. This supports a stronger genetic basis for disease in female-containing families. In simplex families, it has been directly shown that affected females harbor more CNVs and more novel single nucleotide variants (SNVs) affecting autosomal, brain-expressed genes (Jacquemont et al., 2014). However, autism gene discovery has been undertaken primarily in male-only

simplex families, because such families are the most common type. To study the high recurrence risk in affected females, we undertook genetic characterization of autism risk in multiplex families containing two or more female children severely affected with autism, or FEMFs.

Our further requirements for inclusion beyond female sex, that disease be severe and that the families be multiplex, further increases the recurrence risk in such families (Turner et al., 2015). This increased recurrence risk should correspond either to an increased number of autism contributive variants in the FEMF probands or to a greater average risk conferred by each variant present or both. This increased genetic risk in FEMFs is inherited from unaffected parents but its effect is owing to the combination of variants in the affected offspring. In Hirschsprung's disease, another multifactorial neurodevelopmental disorder with a similar male-skewed sex ratio, a positive relationship has been observed between membership in the same three recurrence risk classes represented by FEMFs (female sex, severe disease, and familial disease) and the proportion of individuals with coding variants discovered in the major risk gene (*RET*) (Emison et al., 2010). This supports our assumption that there is a higher burden of more damaging variants in FEMFs probands, and, specifically, that variants affecting the coding sequence, and, therefore, the function of proteins, are more common in FEMFs than in autism families having lower recurrence risk. A preliminary genetic-*cum*-functional study in FEMFs resulted in the identification of *CTNND2*, and the suggestion of *CYFIP1*, as autism risk genes (Turner et al., 2015). This success demonstrates the increased burden of rare coding variation in these families.

Hirschsprung disease, presented in the preceding chapter, represents a case in which we have been able to assign large components of the genetic risk for a multifactorial disorder to several well characterized genes and pathways whose role in the disorder is relatively well understood. Autism spectrum disorder (ASD), on the other hand, represents a case in which there are many known risk genes but where the inherited contributions of these genes to risk for the disorder and the roles of these genes within the disorder are still not fully characterized. This is in part because genetic studies of autism have focused on gene discovery rather than understanding the basis for high familial risk. Many of the largest and most successful gene discovery studies have focused on simplex families, in which familial liability is low. Indeed, such families are expected to be enriched for *de novo* variation contributive to autism, which is relatively easy to identify and interpret (Iossifov et al., 2014; O'Roak et al., 2012; Sanders et al., 2012, 2015; Satterstrom et al., 2020). While these studies have provided valuable insights into genes disrupted in autism, understanding the contribution of *de novo* variation does not immediately lend itself to understanding autism's high familial risk.

Through our study of exome sequencing on 99 such FEMFs, we sought to characterize the contributions of rare coding variation in known and novel autism risk genes within these high recurrence risk families and to determine whether genetic risk in these families is caused by the same or a different set of genes from those that have previously been associated with autism in studies of *de novo* variation.

4.2 Methods

4.2.1 Cohort ascertainment and description

The female-enriched multiplex family (FEMF) cohort we studied consists of 416 exome sequenced individuals from 99 nuclear families with the following characteristics: at least two female children affected with autism spectrum disorder under DSM-5 criteria and at least one affected child (either male or female) having a diagnosis of autistic disorder under DSM-4 on the basis of both the Autism Diagnostic Interview Revised (ADIR) and Autism Diagnostic Observation Schedule (ADOS). The criterion that at least one affected child have a diagnosis of autistic disorder (a more narrowly defined and severe disorder compared to the autism spectrum as a whole) is how we ensured that we were selecting families with severely affected children, characterized by increased recurrence risk. These families were identified through a search of the Autism Genetics Resource Exchange and National Institute of Mental Health Autism Genetics databases. DNA samples were obtained for a total of 420 individuals from 100 families that met the inclusion criteria, though one family was later found to be duplicated and exome sequencing of one individual failed, as detailed in Quality Control.

4.2.2 Exome sequencing methods

Exome capture was performed on the 420 FEMF samples using the Agilent 44Mb Broad Version 2 capture kit, and all libraries were sequenced with 76bp paired-end Illumina sequencing. All sequencing, data preprocessing, and variant calling was performed at the Broad Institute. Variant calling for single nucleotide variants (SNVs)

and small (<50bp) insertions and deletions (Indels) was performed on a combined call-set including the 420 FEMF samples and 371 non-Finnish European ancestry controls (comparable to FEMFs samples in capture and sequencing protocols) drawn from Nation Institute of Mental Health (NIMH) neuropsychiatric controls. Variant calls were made using the GATK Haplotype Caller in GATK version 3.1 (McKenna et al., 2010). Initial filtering of variants was done using the Variant Quality Score Recalibration (VQSR) method within GATK, which is based on detection of known variant sites. For SNVs, HapMap3.3 and Omni2.5 were used as training sites with HapMap3.3 used as the truth set. For VQSR of Indels, a set of curated Indels (Mills_and_1000G_gold_standard.indels.b37.vcf) that comprise the GATK resource bundle were used as both a training and truth set. In filtering of called variants, SNPs and Indel variants up to the 99.9% Truth Sensitivity tranche (0.1% false negative rate) were retained, in order to have sufficient sensitivity to discover novel variants in the FEMF samples. Sequencing failed in one sample leaving 419 samples: averaged across samples, 87% of capture targets were sequenced at a depth of at least 20x.

Only genotypes with PHRED-scaled quality score ≥ 20 and sequencing depth ≥ 10 were used for analyses, except in Mendelian gene filtering, where any genotype based on five or more reads were considered in order to provide more complete information about variant transmission in families where depth was variable across individuals. For enrichment and association testing, variants were only included in analyses if there were less than 20% missing genotypes across cases and 370 NIMH controls.

4.2.3 Variant annotation and classification

Variants in FEMFs and in the 370 controls were primarily annotated using ANNOVAR (Wang et al., 2010). Frequency annotations included frequency in 1000 Genomes super-populations, all Exome Aggregation Consortium (ExAC) subpopulations, and all Exome Variant Server (EVS) subpopulations. Gene annotations used the refGene track downloaded from the UCSC genome browser. Multiple conservation and functional scores were included in annotation of all variants (including PHRED-scaled CADD Scores), but Indels were separately annotated for pathogenicity using SIFT Indel (Hu & Ng, 2013).

For testing the contributions of known genes (identified through *de novo* loss of function in low recurrence risk families) to autism risk in FEMFs, we used two definitions of what constituted a damaging variant: loss-of-function (LOF) variants only and all putatively damaging variants. For both classes of variants, a minor allele frequency (MAF) of ≤ 0.1% was required across all ExAC, EVS, and 1KG samples. LOF variants were defined as nonsense variants and frameshifting Indels predicted by SIFT Indel to be damaging. The more inclusive set of damaging variants also included missense and ± 2 bp splice sites both with PHRED-scaled CADD score ≥ 20 and non-frameshifting indels categorized by SIFT Indel as damaging.

For association-based tests, we used all putatively damaging variants, as defined above, except that we used a more inclusive MAF cutoff of 1% rather than 0.1% used to test previously identified *de novo* risk genes. We used a more stringent MAF threshold

for genes previously identified through *de novo* LOF because we expected true damaging variants in these genes to be highly penetrant in the heterozygous state and sought to exclude less damaging variation from these large and sometimes missense tolerant genes in that analysis. However, for gene discovery in a relatively small cohort, we wished to allow for less penetrant and recessive alleles present in the general population at higher frequency and which we expect to disproportionately contribute to the higher heritable risk in FEMFs.

Prior to our family-based analyses in FEMFs, the GATK genotype refinement workflow was used to identify and exclude variants with apparent *de novo* inheritance in the 77 FEMFs, where both parents had been successfully exome sequenced, as these variants do not contribute to heritable autism and because of the high false positive rate for this class of called variants. This was not done for association-based analyses because such a step could not be taken for available exome sequenced controls.

4.2.4 Sample quality control

First, ancestry and the relatedness between the FEMFs samples were investigated using the SNPRelate package in R (Zheng et al., 2012). For ancestry estimation, 260 samples collected through the 1000 Genomes Project, 10 from each of the 26 populations sampled, were combined with 100 independent FEMFs probands for principal components analysis (PCA) using only autosomal variants with MAF ≥ 5% and missingness ≤ 1% across all samples, linkage disequilibrium (LD) trimmed using an r^2 threshold of 0.2. The 1000 Genomes samples used for ancestry determination were

sequenced and called using the same protocols as the FEMFs samples at the Broad Institute, though they were called separately. Based on plots of the first three principal components (Figure 1), the FEMFs cohort was found to consist of primarily European ancestry families in addition to 9 families with Native American, 4 with African and 3 with East Asian ancestry as well.

As a check on identity and relationships within our cohort, we also used SNPRelate to compute pairwise relatedness coefficients for all individuals in the FEMFs cohort, using the same variant filtering criteria as for PCA. We found that two of the 100 families were, in fact, a single family reported in both databases with different cell ID's, lowering our total number of families to 99. We also identified 5 samples reported as being from affected females that were clearly mislabeled. Four of these samples had no relationships to their reported families and were male (as determined by average depth of coverage on the Y chromosome). The other sample was a duplicate of an unaffected male sibling. Fortunately, there was at least one additional affected individual for each of the families containing mislabeled samples. Additionally, 4 additional sets of monozygotic twins were identified within these families, and 10 that had been reported in the NIMH family information were confirmed. 13 of these sets of identical twins are comprised of concordantly affected females and the other is comprised of concordantly affected males. We removed the 9 anomalous samples to obtain a sequenced FEMFs cohort of 410 samples from 99 families. This includes 187 affected females, 22 affected males, 24 male and female unaffected siblings, 95 mothers, and 82 fathers. Both parents and one affected child were sequenced in 77 of these families.

4.2.5 Recalling genotypes of FEMFs for comparison to SSC

In order to compare rare coding variant-based genetic risk between the FEMF population characterized by high recurrence risk and a collection of low recurrence risk families, we recalled the FEMF exome data and 1,506 mixed ancestry control exomes using the same analysis pipeline used for the SSC (Simons Simplex Collection) Total Recall Project (Krumm et al., 2015), restricting our attention to capture regions common across the different sequencing kits used. Unfortunately, no comparison of burden could be made because of quality and analytic inconsistencies within the SSC exome data across its contributing sequencing centers. To expand somewhat upon these inconsistencies, the number of called variants (both rare and common) in the SSC exome calling data varied widely, in stark contrast to fairly consistent numbers of called variants across FEMFs and control exome calls made with the same calling pipeline. The number of called variants moreover, could not be explained by depth of sequencing, the rate at which sequencing reads could be mapped, or differences in capture efficiency. Interestingly, when SSC sequencing was recalled by us (using what was, based on published methods, the same pipeline, more sites were called, and the data were more consistent. Thus, it is possible that recalling the entire SSC exome dataset would have allowed for comparison to FEMFs. This computational task was not, however, feasible for us, and we proceeded with only our planned analyses that compared FEMFs to collections of controls.

4.2.6 Test of pathogenic variant enrichment in previously reported autism genes

We first wished to confirm the relevance of previously identified autism genes in this high-risk cohort, and, therefore, tested enrichment of deleterious variants within 99 independent probands from each family. For this analysis we used a high confidence set of 28 genes (estimated FDR of 0.01), reported in Sanders et al. 2015, that were drawn from several large-scale autism genomic studies. The total number of LOF and all deleterious variants and the number of genes hit by these two classes of variants within the 28 high confidence autism genes were counted across the 99 probands. FEMF enrichment for variation in these two classes was determined by comparing the number of variants within these two classes in FEMFs to those that occurred in 99 control individuals randomly chosen from the 370 NIMH controls based on 10,000 samples taken with replacement.

4.2.7 Rare variant association testing

In order to detect genes enriched for rare damaging variation in FEMFs, we estimated the probability (P value) of finding as many or a greater number of distinct pathogenic alleles in each gene among 99 FEMF probands, compared to the number of distinct pathogenic alleles (as defined previously) expected on the basis of the frequency of such alleles in 114,704 individuals from the gnomAD database version 2.1.1 (Karczewski et al., 2020) not ascertained on the basis of any neurological disorder. In order to compare the frequency of such alleles between FEMFs and gnomAD, we first

annotated and filter gnomAD variants in the genes we tested as we had previously annotated FEMF variants. However, we did not test all genes. In order to increase our power to find autism genes, we reduced the genes that we tested for association to those where we might reasonably expect loss of function to result in a high fitness effect, which would be expected for highly penetrant autism genes. Therefore, we tested for association in only those genes having a gnomAD pNull intolerance score < 0.5 (i.e., those genes that are classified as having >50% chance of being essential, based on LOF frequency in healthy controls) in order to exclude genes unlikely to have severe, high penetrance phenotypes from consideration. We further restricted our attention to only those genes harboring $\geq$1 pathogenic allele in both FEMF probands and healthy controls in the gnomAD database so that we only tested genes for which we could reasonably estimate the frequency of such variants in both groups. In order to exclude spurious associations resulting from batch-specific calling artifacts, we also excluded genes that had two of more pathogenic alleles in 370 jointly called European ancestry controls. A total of 2,204 genes met all of the criteria set out above and were tested for association.

As the first stage in this association test, we ascertained the observed number of distinct pathogenic alleles (d_0) in the 99 probands (a random variable d). We then estimated the expected number of distinct alleles by simulating 99 diploid genotypes for each gene, assigning gnomAD alleles to individuals on the basis of their allele frequency. We did this with 10,000 replications, using the mean ($\vec{d}$) as the expected rate of distinct alleles. We determined the significance value (α) of the hypothesis of no gene effect as

α = Prob $\{d \geq do | \vec{d}\}$, assuming d to be Poisson distributed. We used a Bonferroni correction to account for the 2,204 tests performed such that we considered only genes having α values at or below 2.3×10^{-5} (equivalent to a threshold of 0.05) to represent significant associations.

4.2.8 Family-based filtering to identify high penetrance genes

In order to identify potential cases of high penetrance recessive inheritance in FEMFs, pathogenic alleles were defined as they were for rare variant association testing. Genes were considered to be potentially recessively inherited within a family if all affected family members had homozygous or compound heterozygous genotypes and all unaffected family members had at least one reference allele.

In order to identify potential cases of incompletely penetrant dominant inheritance in FEMFs, we established a stricter definition of allele pathogenicity. First, in order to exclude variant-calling artifacts, we only retained alleles observed no more than once in 370 jointly called NIMH controls. Second, we used a more stringent MAF cutoff of 0.1% across control databases compared to the 1% cutoff used for recessive gene filtering. Third, we used a more stringent PHRED-scaled CADD score threshold of $\geq$29 to further restrict variants to those that are expected to be the most damaging. In order to consider only those genes expected to have dominant disease inheritance, we also restricted genes to those having a pLI > 0.9, indicating intolerance of single copy loss of function with at greater than 90% probability on the basis of observed genotypes in gnomAD.

In order to test the significance of our dominant and recessive gene findings in FEMFs, we counted instances of potential dominant or recessive inheritance (counting each apparent dominant or recessive gene separately even if it occurred in the same family as another gene), and we used a computer simulation-based approach to determine whether there was enrichment in FEMFs compared to what we would expect, given the parental genotypes for putatively pathogenic alleles. For this test it was necessary to restrict our comparison to the 77 families for which both parental genotypes were available. We also performed a separate test restricted to only the 73 families for which at least two affecteds had been sequenced.

4.3 Results

4.3.1 Enrichment in previously reported autism genes

For both LOF and all deleterious variants, significant FEMF enrichment was observed for the number of genes hit (P = 0.029 for LOF and P = 0.012 for deleterious variants, respectively) but not for the total number of variants. These results are shown in Figure 2. While some of the genes identified as autism genes on the basis of *de novo* association are depleted for variation in controls, others among these genes harbor a great number of rare damaging alleles. Thus, both FEMF probands and controls may harbor deleterious variants driven by a few genes, but variation in less tolerant genes in FEMFs drive enrichment in the number of genes affected.

4.3.2 Rare variant association testing

Among the 2,204 genes tested for rare variant association in FEMFs, none were significantly associated with autism after multiple testing correction. Association information on the top 50 associated genes can be found in Table 1, alongside information on the transmission of variants in those genes to additional affected individuals within FEMFs (beyond the probands, on whom the association is based). 46 of the top 50 associated genes have 2 or more distinct pathogenic alleles in FEMFs, and many are supported by multiple transmissions to affected siblings, providing a basis for stratification of autism candidate genes.

4.3.3 Family-based filtering

After filtering families to identify those where autism might be explained by recessive inheritance, there were seven genes that showed potential compound heterozygous inheritance in eight families, of which five families have two sequenced parents. No families had putatively damaging variants inherited in a manner consistent with homozygous recessive inheritance.

After filtering to identify families where autism might be explained by incompletely penetrant dominant inheritance, we found at least one candidate dominant gene in 42 of the 77 families (with a total of 68 instances) in which both parents and at least one affected child had been exome sequenced. These findings are summarized in Table 3.

In our computer simulations to determine whether putatively damaging parental alleles were transmitted in a manner consistent with recessive or dominant

inheritance more than would be expected by chance, damaging alleles were not transmitted to affected children significantly more than would be expected by chance (Table 4).

4.4 Discussion

The question whether genes discovered through *de novo* mutation make contributions to autism in families characterized by high heritable risk for autism is an important one. It is necessary to know the extent to which variation in these genes contribute to genetic risk in different family types in order to use them meaningfully in counselling high risk families. There is also a question as to whether the same variation that causes autism in males also causes autism in females – it may be that the reason fewer females are affected by autism is that females are only susceptible to certain autism genes. This study indicates that at least some of the 28 genes discovered through *de novo* mutation in primarily male containing families do make a significant contribution to autism genetic risk in high risk female containing families (Sanders et al., 2015). Larger studies that are capable of comparing the impact of these *de novo* risk genes between female and male containing families will ultimately be necessary to quantify the extent to which these genes make contributions. We were unable to attempt this for lack of comparable datasets for individuals at lower risk, but the high rate at which autism families are now being whole genome sequenced will allow these questions, necessary for accurate determination of familial risk, to be answered.

While we failed to detect significant autism associated genes in FEMFs owing to, presumably, the relatively small size of the FEMF cohort and the complexity of autism genetics, it is quite possible that many of the genes most associated with autism in our study contribute to the phenotypes observed in these families, but, in the absence of the ability to functionally validate these genes or to confirm their association in an independent dataset, it is not possible to know for certain. Our previous success in identifying *bona fide* risk genes using association in a relatively small cohort combined with functional validation both in Chapter 3 of this dissertation and in previous studies of FEMFs (Turner et al., 2015) shows the value of sub-significant associations paired with functional validation. This is not surprising given that most genes contributing to a disease do so in only a small fraction of families. The advantage of looking at a smaller cohort like FEMFs with extreme heritable risk compared to identifying autism genes in a larger cohort at low genetic risk is that the discovery of genes is likely to be of greater benefit to affected families, and genes discovered may better contribute to our understanding of the high heritability of autism. Therefore, studies in cohorts such as this one are of great value, though success requires functional assays and independent validation.

There are many well characterized “Mendelian” disorders that contribute to autism risk, and, as we have discussed in previous chapters, these disorders have variable phenotypic features and penetrance. Betancur (2011) identified more than one hundred syndromic disorders with which autism is associated. While we were not able to make any definitive genetic diagnoses in FEMFs using a Mendelian filtering approach,

developmental disorders with autism associations are a potentially fruitful area to investigate in order to understand autism in some families. Many of these known genes have accompanying syndromic features, some of which may be subtle and others not universal (canonical). We did not have access to detailed phenotyping of the individuals in the FEMF cohort, as we did for families described in Chapter 2 of this thesis, and this was likely contributory. Given that we found many potential instances of Mendelian inheritance in these families, it is well worth the additional effort in collecting families to achieve the kind of phenotypic depth that would allow both genetic and phenotypic associations to be used to understand the genes underlying autism or other neurodevelopmental disorders.

4.5 Support and Acknowledgements

Supported by

Simons Foundation for Autism Research grant 309779, funding the exome sequencing of FEMFs

Acknowledgments

We thank Mark Daly and Christine Stevens for overseeing exome sequencing of FEMF samples at the Broad and for their help in navigating additional autism sequencing data. We also thank Nik Krumm and Evan Eichler for their help in interpretation of data from the Simons Simplex Collection Total Recall Project.

4.5 Chapter 4 Tables:

Table 1: *Top 50 autism associated genes from rare variant testing (no significant hits)*

[a]Counts for 99 probands, 370 NIMH controls, and the gnomAD mean all refer to the number of distinct pathogenic alleles. [b]Counts of genotyped affected siblings of proband are totaled across the different proband carriers of damaging variants in each gene and "affected sibling carriers" indicates the number of those additional genotyped affecteds harboring the same pathogenic allele as their proband sibling.

Chromosome	Gene	In 99 probands[a]	In 370 NIMH controls[a]	gnomAD p-value	gnomAD mean[a]	Genotyped affected siblings of proband carriers[b]	Affected sibling carriers[b]
X	*NAA10*	2	0	9.16E-05	0.01	1	0
2	*HNRNPA3*	2	0	3.50E-04	0.03	2	1
12	*SELPLG*	2	1	3.96E-04	0.03	4	0
2	*STAT4*	5	2	8.21E-04	0.71	6	2
10	*PIK3AP1*	3	2	8.35E-04	0.18	3	1
20	*EIF2S2*	2	0	1.00E-03	0.05	2	2
X	*MAP7D3*	2	2	1.27E-03	0.05	2	0
9	*TTF1*	2	0	1.31E-03	0.05	3	1
13	*ZDHHC20*	2	0	1.49E-03	0.06	2	1
2	*PIGF*	2	1	1.59E-03	0.06	1	0
5	*SLC12A2*	3	0	1.78E-03	0.23	3	1
12	*CD27*	2	1	1.87E-03	0.06	2	0
6	*ZKSCAN8*	2	1	1.99E-03	0.06	2	0
X	*DIAPH2*	2	2	1.99E-03	0.06	1	1
12	*TBC1D15*	3	0	2.19E-03	0.25	4	1
2	*SCN3A*	4	0	2.22E-03	0.53	3	3
4	*CCNI*	2	0	2.39E-03	0.07	0	0
9	*LINGO2*	2	0	2.47E-03	0.07	2	1
6	*HLA-A*	2	2	2.50E-03	0.07	1	0
15	*BMF*	2	2	2.62E-03	0.07	3	1
17	*RHBDL3*	3	1	3.08E-03	0.28	4	3
6	*DLK2*	3	1	3.13E-03	0.29	5	3
15	*ZNF280D*	2	1	3.28E-03	0.08	2	1
12	*LGR5*	4	0	4.05E-03	0.63	5	0
5	*IK*	2	1	4.20E-03	0.09	4	0
4	*C4orf27*	2	2	4.22E-03	0.09	2	0
6	*BCLAF1*	2	2	4.25E-03	0.10	3	2
10	*ATOH7*	2	0	4.39E-03	0.10	4	3

19	*LIM2*	2	0	4.42E-03	0.10	2	2
3	*ZNF502*	2	1	4.50E-03	0.10	3	2
22	*TIMP3*	2	1	4.66E-03	0.10	3	2
1	*HIPK1*	3	0	4.71E-03	0.33	4	4
1	*LEPROT*	2	1	4.77E-03	0.10	3	0
X	*TNMD*	2	1	4.82E-03	0.10	1	1
21	*C2CD2*	3	0	4.97E-03	0.34	4	3
14	*SNW1*	2	0	5.30E-03	0.11	3	1
17	*IKZF3*	2	0	5.52E-03	0.11	3	3
8	*PARP10*	3	0	6.42E-03	0.37	2	1
12	*SRRM4*	2	1	6.81E-03	0.12	2	1
15	*SLC27A2*	3	1	7.29E-03	0.39	5	3
5	*CHD1*	3	1	7.57E-03	0.39	5	4
6	*CDSN*	1	0	7.77E-03	0.01	1	0
2	*PPP3R1*	1	2	8.07E-03	0.01	2	0
2	*RAPGEF4*	4	2	8.08E-03	0.77	5	2
3	*CRYGS*	2	1	8.47E-03	0.14	1	1
1	*STRIP1*	3	2	8.50E-03	0.41	4	2
9	*GDA*	2	0	8.52E-03	0.14	2	1
4	*OCIAD2*	1	1	8.66E-03	0.01	1	0
3	*LHFPL4*	1	0	8.86E-03	0.01	2	0
1	HMGN2	1	0	8.96E-03	0.01	1	0

Table 2: *Instances of compound heterozygous inheritance*

Variants and genotypes for FEMFs are given with respect to the amino acid sequence change. Families indicated with an * have only one genotyped parent. [a]These two variants are inherited on the same haplotype. [b] "-" indicates that the variant was not present in the database.

Family	Gene	Transcript	Variant	Father	Mother	Proband #1	Proband #2	Unaffected #1	CADD	ExAC MAF[b]
50002	*EPPK1*	NM_031308	G2270S	G/G	G/S[a]	ambiguous	G/S	G/G	36	-
			A2259S	A/A	A/S[a]	A/S	A/S	A/A	20.8	0.0009
			A1900V	A/V	A/A	A/V	A/V	A/V	33	0.001
50004	*LRBA*	NM_001199282	N815S	N/N	N/S	N/S	N/S	unknown	24.7	0.0018
			F598C	F/C	F/F	F/C	F/C	unknown	25.4	-
50016	*PRDM2*	NM_001007257	S1158N	S/N	S/S	S/N	S/N	-	21.8	-
			1056del	E/E	E/-	E/-	E/-	-	NA	0.0041
50032*	*PREX1*	NM_020820	R1589K	R/R	unknown	R/K	unknown	unknown	31	0.0009
			D794N	D/N	unknown	D/N	unknown	unknown	21	0.0008
50044	*KIAA0430*	NM_001184998	S1620F	S/F	S/S	S/F	S/F	-	34	0.0042
			G4R	G/G	G/R	G/R	G/R	-	28	0.0041
50058*	*EPPK1*	NM_031308	E2295K	E/E	unknown	E/K	E/K	E/E	35	0.0002
			V1902fs	V/fs	unknown	V/fs	V/fs	V/fs	NA	0.0006
50078	*SLC16A5*	NM_001271765	S343T	S/S	S/T	S/T	unknown	S/S	22.7	0.00005511
			37del	F/-	F/F	F/-	unknown	F/-	NA	0.0001
50093*	*FBXO25*	NM_012173	D163H	unknown	D/D	D/H	D/H	-	21.8	0.0002
			K290M	unknown	K/M	K/M	K/M	-	21.1	-

Table 3: *Instances of possible dominant inheritance in FEMFs*

As in Table 2, variants and genotypes for FEMFs are given with respect to the amino acid sequence change. Note that there are several instances in which dominant multigenic inheritance is a possibility. [a]This is a start-loss Indel variant.

Family	Gene	Transcript	Variant	Father	Mother	Proband #1	Proband #2	Other affected #1	Other affected #2	Unaffected #1	Unaffected #2	CADD	ExAC MAF
50000	*ATAD2*	NM_014109	I527V	I/V	I/I	I/V	unknown	-	-	-	-	31	1.17E-05
	ATRN	NM_001207047	A1063T	A/T	A/A	A/T	unknown	-	-	-	-	35	1.10E-05
	FBN1	NM_000138	G1482S	G/S	G/G	G/S	unknown	-	-	-	-	36	2.20E-05
50004	*JARID2*	NM_001267040	R655Q	R/Q	R/R	R/Q	R/Q	-	-	unknown	-	36	-
	TULP4	NM_020245	A1516V	A/V	A/A	A/V	A/V	-	-	unknown	-	33	0.0003
50005	*MAP3K4*	NM_001301072	D279N	D/N	D/D	D/N	D/N	-	-	-	-	34	4.44E-05
50007	*IPO7*	NM_006391	R252Q	R/R	R/Q	R/Q	R/Q	-	-	R/R	-	31	2.21E-05
	SIK3	NM_001281749	V173I	V/I	V/V	V/I	V/I	-	-	V/V	-	32	-
	TSC2	NM_001318831	R439Q	R/Q	R/R	R/Q	R/Q	-	-	R/R	-	34	4.48E-05
50013	*HEATR1*	NM_018072	S28C	S/S	S/C	-	S/C	-	-	-	-	33	-
50016	*HEATR1*	NM_018072	P1397S	P/P	P/S	P/S	P/S	-	-	-	-	31	3.33E-05
50018	*ADNP*	NM_001282532	A1017Gfs	A/fs	A/A	-	A/fs	-	-	-	-	NA	1.11E-05
	EIF3G	NM_003755	G76S	G/S	G/G	-	G/S	-	-	-	-	33	-
	MPRIP	NM_001364716	D562N	D/D	D/N	-	D/N	-	-	-	-	35	3.31E-05
50020	*CLUH*	NM_001366661	R1067C	R/R	R/C	R/C	R/C	-	-	unknown	-	29.1	1.26E-05
	NMT1	NM_021079	G205S	G/G	G/S	G/S	G/S	-	-	unknown	-	32	5.51E-05
	PPP1R13B	NM_015316	R380I	R/I	R/R	R/I	R/I	-	-	unknown	-	32	0.0002
50024	*TLN1*	NM_006289	A2013T	A/T	A/A	A/T	A/T	A/T	A/T	unknown	unknown	31	0.0005
50025	*HMGCS1*	NM_001324219	V162A	V/V	V/A	-	V/A	V/A	-	V/V	-	29.6	2.21E-05
50027	*HERC2*	NM_004667	G2974R	G/G	G/R	G/R	unknown	-	-	-	-	29.1	-
	HGF	NM_000601	G186E	G/G	G/E	G/E	unknown	-	-	-	-	33	2.21E-05
50029	*DMBX1*	NM_147192	T89N	T/N	T/T	T/N	T/N	T/N	unknown	-	-	33	-

50030	*BICD2*	NM_001003800	E157Rfs	E/fs	E/E	E/fs	E/fs	E/fs	-	-	-	NA	-
50036	*GMPS*	NM_003875	P612L	P/P	P/L	P/L	P/L	-	-	P/P	-	34	-
	KDM4A	NM_014663	K301R	K/K	K/R	K/R	K/R	-	-	K/K	-	35	1.10E-05
50037	*ARIH2*	NM_001317334	V37M	V/M	V/V	V/M	V/M	-	-	V/V	-	29.2	1.10E-05
50041	*KIAA0368*	NM_001363756	V1726E	V/V	V/E	V/E	V/E	-	-	V/V	V/V	32	-
	LRP12	NM_001135703	S598X	S/X	S/S	S/X	S/X	-	-	S/S	S/S	39	-
50042	*CLOCK*	NM_004898	Q626del	Q/del	Q/Q	Q/del	Q/del	Q/del	-	-	-	NA	6.62E-05
	TSHZ1	NM_001308210	P342Afs	P/fs	P/P	P/fs	P/fs	P/fs	-	-	-	NA	-
50043	*SNX2*	NM_003100	F74L	F/F	F/L	F/L	F/L	unknown	-	F/F	-	33	-
50044	*CAMK2G*	NM_001367524	G14W	G/G	G/W	G/W	G/W	unknown	-	-	-	31	1.10E-05
50046	*CSNK1D*	NM_001363749	S114Vfs	S/S	S/fs	S/fs	S/fs	-	-	-	-	NA	-
50053	*MKL1*	NM_001282660	E126K	E/K	E/E	E/K	E/K	-	-	-	-	37	1.10E-05
50059	*HNRNPA3*	NM_001330247	K112del	K/K	K/del	K/del	-	-	-	-	-	NA	-
50062	*WDFY3*	NM_014991	R3166X	R/X	R/R	R/X	R/X	unknown	-	-	-	52	-
50064	*CLK2*	NM_001294339	R267Q	R/Q	R/R	R/Q	R/Q	R/Q	-	unknown	unknown	34	-
50065	*CHD2*	NM_001271	R1678Q	R/Q	R/R	R/Q	R/Q	-	-	-	-	33	9.92E-05
	DOCK9	NM_001130049	K715E	K/K	K/E	K/E	K/E	-	-	-	-	29.9	1.12E-05
50067	*TNRC6B*	NM_015088	S1537G	S/G	S/S	S/G	S/G	-	-	S/S	-	32	-
50068	*AHCTF1*	NM_001323342	R209H	R/R	R/H	R/H	R/H	-	-	-	-	32	1.10E-05
50071	*C6orf136*	NM_001109938	P47L	P/L	P/P	P/L	P/L	-	-	unknown	unknown	37	-
	DYRK1A	NM_001347723	R14C	R/R	R/C	R/C	R/C	-	-	unknown	unknown	32	4.41E-05
	NR6A1	NM_001278546	S181L	S/S	S/L	S/L	S/L	-	-	unknown	unknown	36	-
	SYT11	NM_152280	S63del	S/del	S/S	S/del	S/del	-	-	unknown	unknown	NA	-
50072	*COPB1*	NM_001144061	E464V	E/E	E/V	E/V	E/V	-	-	unknown	-	33	-
50076	*ANKRD11*	NM_001256183	K1012del	K/del	K/K	K/del	K/del	-	-	-	-	NA	8.82E-05
	MYH9	NM_002473	E1225K	E/E	E/K	E/K	E/K	-	-	-	-	35	1.10E-05
	TMEM201	NM_001010866	P85L	P/P	P/L	P/L	P/L	-	-	-	-	29	2.32E-05
	XYLT1	NM_022166	R754H	R/R	R/H	R/H	R/H	-	-	-	-	34	7.76E-05

50077	*APC2*	NM_001351273	E189K	E/K	E/E	E/K	E/K	-	-	unknown	-	36	1.55E-05
50078	*EFNB2*	NM_004093	N282del	N/del	N/N	N/del	unknown	-	-	N/N	-	NA	0.0001
	TRIO	NM_007118	E2656K	E/K	E/E	E/K	unknown	-	-	E/E	-	35	1.10E-05
50080	*MLXIP*	NM_014938	P201L	P/L	P/P	P/L	P/L	-	-	-	-	35	0.0001
50083	*CTPS1*	NM_001905	A129V	A/A	A/V	A/V	A/V	-	-	-	-	34	0.0001
	IGF2R	NM_000876	R1325H	R/H	R/R	R/H	R/H	-	-	-	-	33	1.10E-05
50085	*TRRAP*	NM_003496	K1839M	K/M	K/K	K/M	K/M	K/M	-	-	-	30	-
50089	*DOPEY1*	NM_001199942	R944H	R/R	R/H	R/H	R/H	-	-	unknown	-	31	0.0001
	NAV2	NM_001111018	S224N	S/S	S/N	S/N	S/N	-	-	unknown	-	32	0.0004
50090	*RELN*	NM_005045	A1568V	A/A	A/V	A/V	A/V	-	-	-	-	32	2.21E-05
50091	*LRP2*	NM_004525	R2181H	R/H	R/R	R/H	R/H	-	-	-	-	33	7.78E-05
50094	*KIF1B*	NM_001365951	A219S	A/A	A/S	A/S	A/S	-	-	unknown	-	36	-
	LAMC1	NM_002293	A333V	A/V	A/A	A/V	A/V	-	-	unknown	-	29.1	1.10E-05
	WNT3A	NM_033131	S181Hfs	S/fs	S/S	S/fs	S/fs	-	-	unknown	-	NA	2.55E-05
50095	*GATAD2B*	NM_020699	E23K	E/E	E/K	E/K	E/K	-	-	unknown	-	30	6.62E-05
50097	*HIPK1*	NM_198269	M1?[a]	M/M	M/?	M/?	M/?	M/?	-	-	-	35	-
50100	*DNAJA2*	NM_005880	R388H	R/R	R/H	R/H	R/H	-	-	unknown	-	33	0.0002
	SLC20A2	NM_001257180	R609C	R/C	R/R	R/C	R/C	-	-	unknown	-	36	0.0001

Table 4: *Test for over-transmission of putatively damaging parental alleles*

Here we compared the number of instances of possible dominant and recessive inheritance to simulations based on random segregation of alleles. [a] Here quads indicates families in which there are two genotyped parents and two genotyped affected siblings. [b]Note that there are two additional instances of possible dominant inheritance used for comparison to simulated counts – beyond those presented in Tables 2 and 3; these two additional instances were removed after the protein annotation was found to be incorrect, but these variants were used for comparison to simulations in order to ensure comparability between the FEMF observed and simulated patterns of inheritance.

Inheritance pattern	Instances in 77 FEMF trios	Instances in 73 FEMF quads[a]	In 10,000 simulations of 77 trios	In 10,000 simulations of 73 quads
Recessive	5	4	p = 0.79 (mean = 6.49, SD = 2.43)	p = 0.74 (mean = 4.96, SD =2.17)
Dominant	70[a]	58[a]	p = 0.62 (mean = 71.86, SD = 7.32)	p = 0.70 (mean = 61.38, SD = 6.94)

4.6 Chapter 4 Figures:

Figure 1: *Determining ancestry of FEMFs through PCA*

FEMFs are plotted alongside NIMH controls and 1KG African ancestry (AFR), Native American admixed ancestry (AMR), East Asian ancestry (ASN), European Ancestry (EUR) and South Asian Ancestry (SAN) controls.

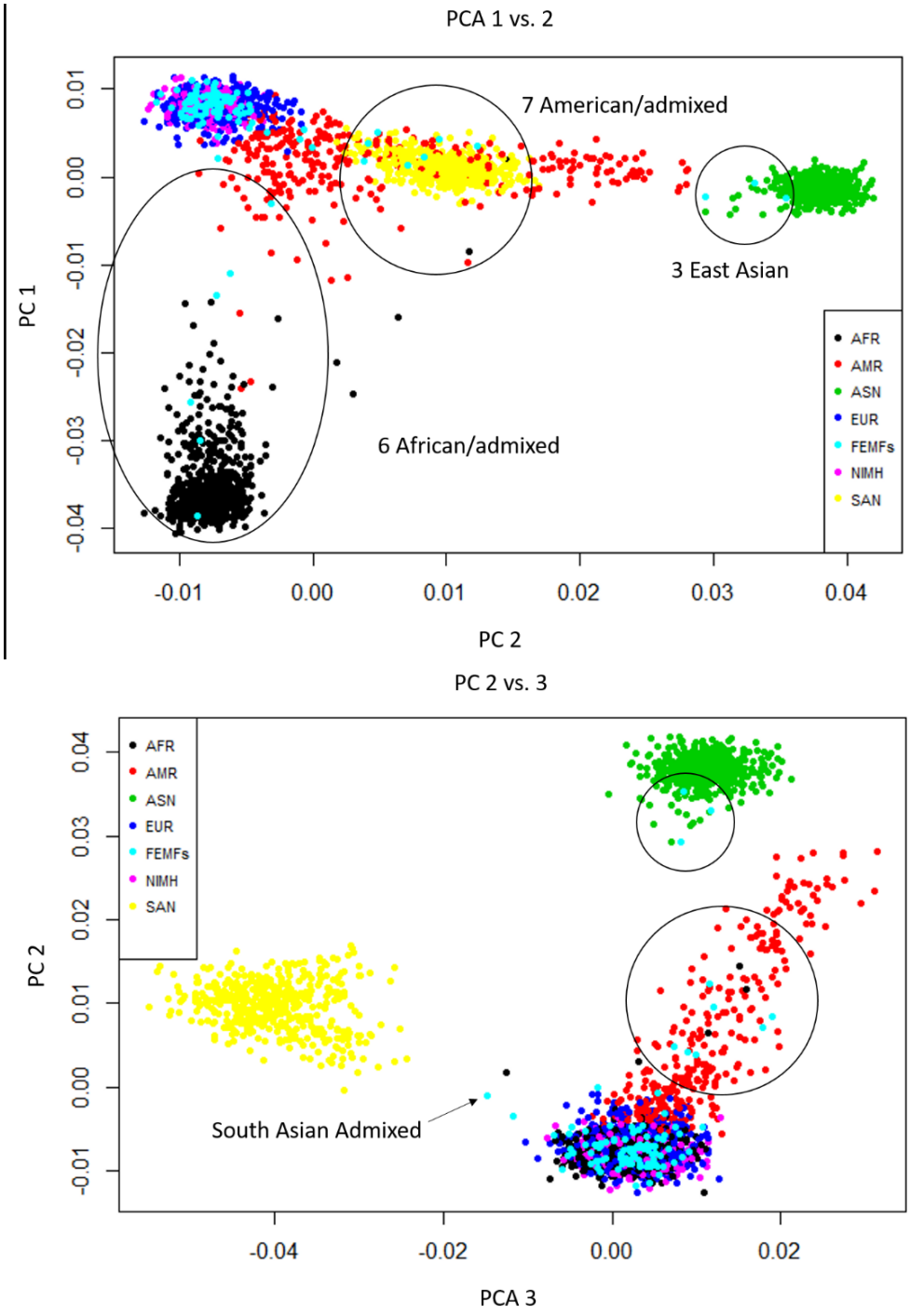
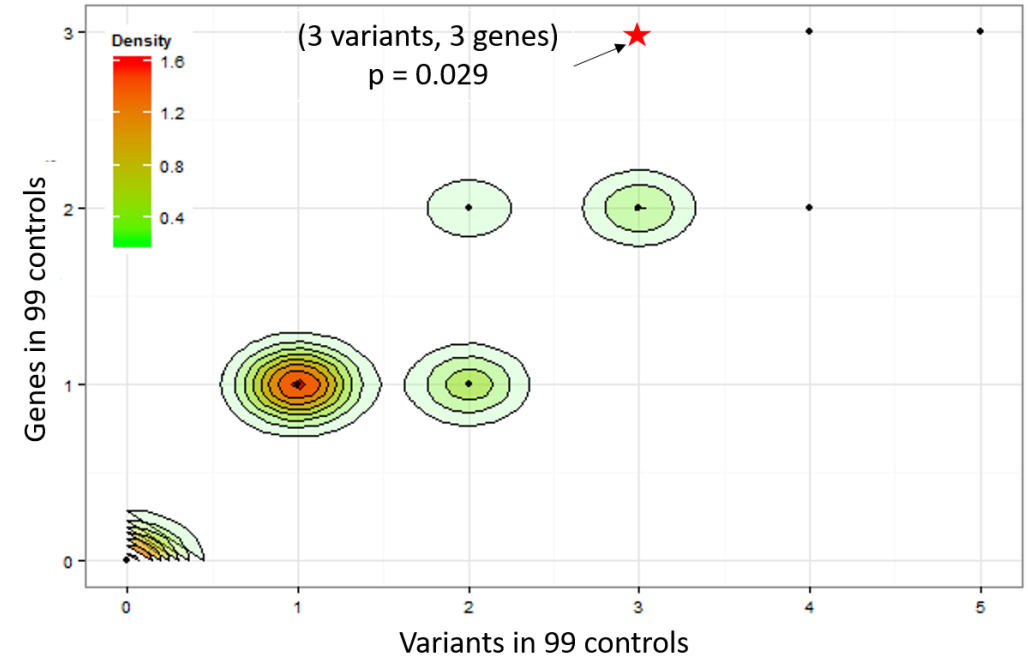
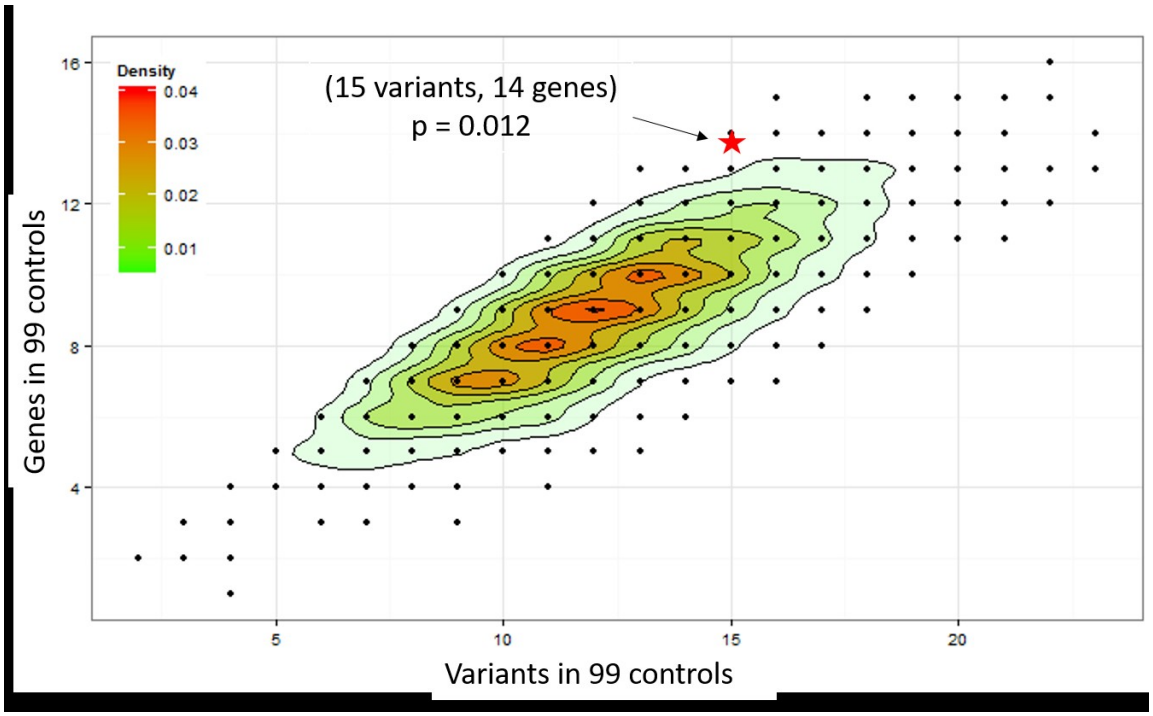


Figure 2: *FEMFs have rare damaging variants affecting a greater than expected number of genes previously associated with autism*

The number of genes and distinct variants affecting those genes are plotted for 10,000 individuals sampled from 370 NIMH controls with replacement, with the genes and distinct variants observed in FEMFs indicated by red stars.





Bibliography

Abràmoff, M. D., Magalhães, P. J., & Ram, S. J. (2004). Image Processing with ImageJ. *Biophotonics International*, *11*, 36–42.

Almeida, A. M., Murakami, Y., Layton, D. M., Hillmen, P., Sellick, G. S., Maeda, Y., Richards, S., Patterson, S., Kotsianidis, I., Mollica, L., Crawford, D. H., Baker, A., Ferguson, M., Roberts, I., Houlston, R., Kinoshita, T., & Karadimitris, A. (2006). Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. *Nature Medicine*, *12*(7), 846–851. https://doi.org/10.1038/nm1410

Amiel, J., Sproat-Emison, E., Garcia-Barcelo, M., Lantieri, F., Burzynski, G., Borrego, S., Pelet, A., Arnold, S., Miao, X., Griseri, P., Brooks, A. S., Antinolo, G., de Pontual, L., Clement-Ziza, M., Munnich, A., Kashuk, C., West, K., Wong, K. K.-Y., Lyonnet, S., … for the Hirschsprung Disease Consortium. (2008). Hirschsprung disease, associated syndromes and genetics: A review. *Journal of Medical Genetics*, *45*(1), 1–14. https://doi.org/10.1136/jmg.2007.053959

Angrist, M., Bolk, S., Halushka, M., Lapchak, P. A., & Chakravarti, A. (1996). Germline mutations in glial cell line-derived neurotrophic factor (GDNF) and RET in a Hirschsprung disease patient. *Nature Genetics*, *14*(3), 341–344. https://doi.org/10.1038/ng1196-341

Arnold, S., Pelet, A., Amiel, J., Borrego, S., Hofstra, R., Tam, P., Ceccherini, I., Lyonnet, S., Sherman, S., & Chakravarti, A. (2009). Interaction between a chromosome 10 RET enhancer and chromosome 21 in the Down syndrome–Hirschsprung disease association. *Human Mutation*, *30*(5), 771–775. https://doi.org/10.1002/humu.20944

Azeem, Z., Naqvi, S. K.-U.-H., Ansar, M., Wali, A., Naveed, A. K., Ali, G., Hassan, M. J., Tariq, M., Basit, S., & Ahmad, W. (2009). Recurrent mutations in functionally-related EDA and EDAR genes underlie X-linked isolated hypodontia and autosomal recessive hypohidrotic ectodermal dysplasia. *Archives of Dermatological Research*, *301*(8), 625–629. https://doi.org/10.1007/s00403-009-0975-1

Badner, J. A., Sieber, W. K., Garver, K. L., & Chakravarti, A. (1990). A genetic study of Hirschsprung disease. *American Journal of Human Genetics*, *46*(3), 568–580.

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., Adam, S., Bulman, D. E., Jones, S. J. M., Avard, D., Nguyen, M. T., Rousseau, F., Marshall, C., Wintle, R. F., Shen, Y., … Boycott, K. M. (2014). FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project. *The*

American Journal of Human Genetics, *94*(6), 809–817. https://doi.org/10.1016/j.ajhg.2014.05.003

Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research*, *1380*, 42–77. https://doi.org/10.1016/j.brainres.2010.11.078

Bodian, M., & Carter, O. O. (1963). A family study of Hirschsprung's disease. *Annals of Human Genetics*, *26*(3), 261–277. https://doi.org/10.1111/j.1469-1809.1963.tb01983.x

Bonanni, P., Negrin, S., Volzone, A., Zanotta, N., Epifanio, R., Zucca, C., Osanni, E., Petacchi, E., & Fabbro, F. (2017). Electrical status epilepticus during sleep in Mowat–Wilson syndrome. *Brain and Development*, *39*(9), 727–734. https://doi.org/10.1016/j.braindev.2017.04.013

Brady, P. D., Moerman, P., De Catte, L., Deprest, J., Devriendt, K., & Vermeesch, J. R. (2014). Exome sequencing identifies a recessive PIGN splice site mutation as a cause of syndromic Congenital Diaphragmatic Hernia. *European Journal of Medical Genetics*, *57*(9), 487–493. https://doi.org/10.1016/j.ejmg.2014.05.001

Carrasquillo, M. M., McCallion, A. S., Puffenberger, E. G., Kashuk, C. S., Nouri, N., & Chakravarti, A. (2002). Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature Genetics*, *32*(2), 237–244. https://doi.org/10.1038/ng998

Carter, C. O. (1969). Genetics of Common Disorders. *British Medical Bulletin*, *25*(1), 52–57. https://doi.org/10.1093/oxfordjournals.bmb.a070671

Chakravarti, A. (2011). Genomic contributions to Mendelian disease. *Genome Research*, *21*(5), 643–644. https://doi.org/10.1101/gr.123554.111

Chakravarti, A., & Lyonnet, S. (2001). Hirschsprung disease. In C. Scriver, A. Beaudet, D. Valle, & et al. (Eds.), *The metabolic & molecular bases of inherited disease* (8th ed, pp. 6231–6255). McGraw-Hill. https://catalyst.library.jhu.edu/catalog/bib_2180131

Chatterjee, S., Kapoor, A., Akiyama, J. A., Auer, D. R., Lee, D., Gabriel, S., Berrios, C., Pennacchio, L. A., & Chakravarti, A. (2016). Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease. *Cell*, *167*(2), 355-368.e10. https://doi.org/10.1016/j.cell.2016.09.005

Cheng, H.-Q., Huang, E.-M., Xu, M.-Y., Shu, S.-Y., & Tang, S.-J. (2012). PVRL1 as a Candidate Gene for Nonsyndromic Cleft Lip With or Without Cleft Palate: No Evidence for the Involvement of Common or Rare Variants in Southern Han

Chinese Patients. *DNA and Cell Biology*, *31*(7), 1321–1327. https://doi.org/10.1089/dna.2011.1556

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., McMillin, M. J., Wiszniewski, W., Gambin, T., Coban Akdemir, Z. H., Doheny, K., Scott, A. F., Avramopoulos, D., Chakravarti, A., Hoover-Fong, J., Mathews, D., Witmer, P. D., Ling, H., … Bamshad, M. J. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics*, *97*(2), 199–215. https://doi.org/10.1016/j.ajhg.2015.06.009

Cluzeau, C., Hadj-Rabia, S., Jambou, M., Mansour, S., Guigue, P., Masmoudi, S., Bal, E., Chassaing, N., Vincent, M.-C., Viot, G., Clauss, F., Manière, M.-C., Toupenay, S., Le Merrer, M., Lyonnet, S., Cormier-Daire, V., Amiel, J., Faivre, L., de Prost, Y., … Smahi, A. (2011). Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases. *Human Mutation*, *32*(1), 70–72. https://doi.org/10.1002/humu.21384

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W. M., Vulto-van Silfhout, A. T., Bosco, P., Friend, K. L., Baker, C., Buono, S., Vissers, L. E. L. M., Schuurs-Hoeijmakers, J. H., Hoischen, A., Pfundt, R., Krumm, N., Carvill, G. L., Li, D., Amaral, D., Brown, N., Lockhart, P. J., … Eichler, E. E. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature Genetics*, *46*(10), 1063–1071. https://doi.org/10.1038/ng.3092

Cuny, H., Rapadas, M., Gereis, J., Martin, E. M. M. A., Kirk, R. B., Shi, H., & Dunwoodie, S. L. (2020). NAD deficiency due to environmental factors or gene–environment interactions causes congenital malformations and miscarriage in mice. *Proceedings of the National Academy of Sciences*, *117*(7), 3738–3747. https://doi.org/10.1073/pnas.1916588117

Dasgupta, R., & Langer, J. C. (2008). Evaluation and Management of Persistent Problems After Surgery for Hirschsprung Disease in a Child. *Journal of Pediatric Gastroenterology and Nutrition*, *46*(1), 13–19. https://doi.org/10.1097/01.mpg.0000304448.69305.28

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Drumm, M. L., Konstan, M. W., Schluchter, M. D., Handler, A., Pace, R., Zou, F., Zariwala, M., Fargo, D., Xu, A., Dunn, J. M., Darrah, R. J., Dorfman, R., Sandford, A. J.,

Corey, M., Zielenski, J., Durie, P., Goddard, K., Yankaskas, J. R., Wright, F. A., & Knowles, M. R. (2005). Genetic Modifiers of Lung Disease in Cystic Fibrosis. *New England Journal of Medicine*, *353*(14), 1443–1453. https://doi.org/10.1056/NEJMoa051469

Eaton, A., Hartley, T., Kernohan, K., Ito, Y., Lamont, R., Parboosingh, J., Barrowman, N., Innes, A. M., & Boycott, K. (2020). When to think outside the autozygome: Best practices for exome sequencing in "consanguineous" families. *Clinical Genetics*, *97*(6), 835–843. https://doi.org/10.1111/cge.13736

Edery, P., Lyonnet, S., Mulligan, L. M., Pelet, A., Dow, E., Abel, L., Holder, S., Nihoul-Fékété, C., Ponder, B. A. J., & Munnich, A. (1994). Mutations of the RET proto-oncogene in Hirschsprung's disease. *Nature*, *367*(6461), 378–380. https://doi.org/10.1038/367378a0

Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: Variant prioritization and Mendelian disease. *Nature Reviews Genetics*, *18*(10), 599–612. https://doi.org/10.1038/nrg.2017.52

Emison, E. S., Garcia-Barcelo, M., Grice, E. A., Lantieri, F., Amiel, J., Burzynski, G., Fernandez, R. M., Hao, L., Kashuk, C., West, K., Miao, X., Tam, P. K. H., Griseri, P., Ceccherini, I., Pelet, A., Jannot, A.-S., de Pontual, L., Henrion-Caude, A., Lyonnet, S., … Chakravarti, A. (2010). Differential Contributions of Rare and Common, Coding and Noncoding Ret Mutations to Multifactorial Hirschsprung Disease Liability. *The American Journal of Human Genetics*, *87*(1), 60–74. https://doi.org/10.1016/j.ajhg.2010.06.007

Emison, E. S., McCallion, A. S., Kashuk, C. S., Bush, R. T., Grice, E., Lin, S., Portnoy, M. E., Cutler, D. J., Green, E. D., & Chakravarti, A. (2005). A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, *434*(7035), 857–863. https://doi.org/10.1038/nature03467

Exome Aggregation Consortium, Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Falk Kieri, C., Bergendal, B., Lind, L. K., Schmitt-Egenolf, M., & Stecksén-Blicks, C. (2014). EDAR-induced hypohidrotic ectodermal dysplasia: A clinical study on signs and symptoms in individuals with a heterozygous c.1072C > T mutation. *BMC Medical Genetics*, *15*, 57. https://doi.org/10.1186/1471-2350-15-57

Fombonne, E. (2003). Epidemiological Surveys of Autism and Other Pervasive Developmental Disorders: An Update. *Journal of Autism and Developmental Disorders*, *33*(4), 365–382. https://doi.org/10.1023/A:1025054610557

Freeze, H. H., Eklund, E. A., Ng, B. G., & Patterson, M. C. (2012). Neurology of inherited glycosylation disorders. *The Lancet Neurology*, *11*(5), 453–466. https://doi.org/10.1016/S1474-4422(12)70040-6

Garcia-Barcelo, M.-M., Tang, C. S., Ngan, E. S., Lui, V. C., Chen, Y., So, M., Leon, T. Y., Miao, X., Shum, C. K., Liu, F., Yeung, M., Yuan, Z., Guo, W., Liu, L., Sun, X., Huang, L., Tou, J., Song, Y., Chan, D., … Tam, P. K. (2009). Genome-wide association study identifies *NRG1* as a susceptibility locus for Hirschsprung's disease. *Proceedings of the National Academy of Sciences*, *106*(8), 2694–2699. https://doi.org/10.1073/pnas.0809630105

Ghoumid, J., Drevillon, L., Alavi-Naini, S. M., Bondurand, N., Rio, M., Briand-Suleau, A., Nasser, M., Goodwin, L., Raymond, P., Yanicostas, C., Goossens, M., Lyonnet, S., Mowat, D., Amiel, J., Soussi-Yanicostas, N., & Giurgea, I. (2013). ZEB2 zinc-finger missense mutations lead to hypomorphic alleles and a mild Mowat–Wilson syndrome. *Human Molecular Genetics*, *22*(13), 2652–2661. https://doi.org/10.1093/hmg/ddt114

Gorelenkova Miller, O., & Mieyal, J. J. (2015). Sulfhydryl-mediated redox signaling in inflammation: Role in neurodegenerative diseases. *Archives of Toxicology*, *89*(9), 1439–1467. https://doi.org/10.1007/s00204-015-1496-7

Goyal, M., Pradhan, G., Gupta, S., & Kapoor, S. (2015). Hypohidrotic ectodermal dysplasia with ankylosis of temporomandibular joint and cleft palate: A rare presentation. *Contemporary Clinical Dentistry*, *6*(1), 110–112. https://doi.org/10.4103/0976-237X.149304

Grozeva, D., Carss, K., Spasic-Boskovic, O., Parker, M. J., Archer, H., Firth, H. V., Park, S.-M., Canham, N., Holder, S. E., Wilson, M., Hackett, A., Field, M., Floyd, J. A. B., Hurles, M., & Raymond, F. L. (2014). De Novo Loss-of-Function Mutations in SETD5, Encoding a Methyltransferase in a 3p25 Microdeletion Syndrome Critical Region, Cause Intellectual Disability. *American Journal of Human Genetics*, *94*(4), 618–624. https://doi.org/10.1016/j.ajhg.2014.03.006

Gui, H., Schriemer, D., Cheng, W. W., Chauhan, R. K., Antiňolo, G., Berrios, C., Bleda, M., Brooks, A. S., Brouwer, R. W. W., Burns, A. J., Cherny, S. S., Dopazo, J., Eggen, B. J. L., Griseri, P., Jalloh, B., Le, T.-L., Lui, V. C. H., Luzón-Toro, B., Matera, I., … Hofstra, R. M. W. (2017). Whole exome sequencing coupled with unbiased functional analysis reveals new Hirschsprung disease genes. *Genome Biology*, *18*(1), 48. https://doi.org/10.1186/s13059-017-1174-6

Haldane, B. J. B. S. (1956). The Estimation and Significance of the Logarithm of a Ratio of Frequencies. *Annals of Human Genetics*, *20*(4), 309–311. https://doi.org/10.1111/j.1469-1809.1955.tb01285.x

Hansen, L., Tawamie, H., Murakami, Y., Mang, Y., ur Rehman, S., Buchert, R., Schaffer, S., Muhammad, S., Bak, M., Nöthen, M. M., Bennett, E. P., Maeda, Y., Aigner, M., Reis, A., Kinoshita, T., Tommerup, N., Baig, S. M., & Abou Jamra, R. (2013). Hypomorphic Mutations in PGAP2, Encoding a GPI-Anchor-Remodeling Protein, Cause Autosomal-Recessive Intellectual Disability. *The American Journal of Human Genetics*, *92*(4), 575–583. https://doi.org/10.1016/j.ajhg.2013.03.008

Harrington, D. P., & Fleming, T. R. (1982). A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*, *69*(3), 553–566. https://doi.org/10.2307/2335991

Hartl, D. L., & Campbell, R. B. (1982). Allele multiplicity in simple Mendelian disorders. *American Journal of Human Genetics*, *34*(6), 866–873.

Honegger, K., & de Bivort, B. (2018). Stochasticity, individuality and behavior. *Current Biology*, *28*(1), R8–R12. https://doi.org/10.1016/j.cub.2017.11.058

Hong, Y., Maeda, Y., Watanabe, R., Ohishi, K., Mishkind, M., Riezman, H., & Kinoshita, T. (1999). Pig-n, a Mammalian Homologue of Yeast Mcd4p, Is Involved in Transferring Phosphoethanolamine to the First Mannose of the Glycosylphosphatidylinositol. *Journal of Biological Chemistry*, *274*(49), 35099–35106. https://doi.org/10.1074/jbc.274.49.35099

Horn, D., Krawitz, P., Mannhardt, A., Korenke, G. C., & Meinecke, P. (2011). Hyperphosphatasia-mental retardation syndrome due to PIGV mutations: Expanded clinical spectrum. *American Journal of Medical Genetics Part A*, *155*(8), 1917–1922. https://doi.org/10.1002/ajmg.a.34102

Hu, J., & Ng, P. C. (2013). SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. *PLoS ONE*, *8*(10), e77940. https://doi.org/10.1371/journal.pone.0077940

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., … Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. https://doi.org/10.1038/nature13908

Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R. M., Myers, R. M., Ridker, P. M., Chasman, D. I., Mefford, H., Ying, P., Nickerson, D. A., & Eichler, E. E. (2009). Population Analysis of Large Copy Number Variants and Hotspots of

Human Genetic Disease. *The American Journal of Human Genetics*, *84*(2), 148–161. https://doi.org/10.1016/j.ajhg.2008.12.014

Ivanovski, I., Djuric, O., Caraffi, S. G., Santodirocco, D., Pollazzon, M., Rosato, S., Cordelli, D. M., Abdalla, E., Accorsi, P., Adam, M. P., Ajmone, P. F., Badura-Stronka, M., Baldo, C., Baldi, M., Bayat, A., Bigoni, S., Bonvicini, F., Breckpot, J., Callewaert, B., … Garavelli, L. (2018). Phenotype and genotype of 87 patients with Mowat–Wilson syndrome and recommendations for care. *Genetics in Medicine*, *20*(9), 965–975. https://doi.org/10.1038/gim.2017.221

Iwata, J., Suzuki, A., Yokota, T., Ho, T.-V., Pelikan, R., Urata, M., Sanchez-Lara, P. A., & Chai, Y. (2014). TGFβ regulates epithelial-mesenchymal interactions through WNT signaling activity to control muscle development in the soft palate. *Development (Cambridge, England)*, *141*(4), 909–917. https://doi.org/10.1242/dev.103093

Jacquemont, S., Coe, B. P., Hersch, M., Duyzend, M. H., Krumm, N., Bergmann, S., Beckmann, J. S., Rosenfeld, J. A., & Eichler, E. E. (2014). A Higher Mutational Burden in Females Supports a "Female Protective Model" in Neurodevelopmental Disorders. *American Journal of Human Genetics*, *94*(3), 415–425. https://doi.org/10.1016/j.ajhg.2014.02.001

Jiang, Q., Arnold, S., Heanue, T., Kilambi, K. P., Doan, B., Kapoor, A., Ling, A. Y., Sosa, M. X., Guy, M., Jiang, Q., Burzynski, G., West, K., Bessling, S., Griseri, P., Amiel, J., Fernandez, R. M., Verheij, J. B. G. M., Hofstra, R. M. W., Borrego, S., … Chakravarti, A. (2015). Functional Loss of Semaphorin 3C and/or Semaphorin 3D and Their Epistatic Interaction with Ret Are Critical to Hirschsprung Disease Liability. *The American Journal of Human Genetics*, *96*(4), 581–596. https://doi.org/10.1016/j.ajhg.2015.02.014

Johnston, J. J., Gropman, A. L., Sapp, J. C., Teer, J. K., Martin, J. M., Liu, C. F., Yuan, X., Ye, Z., Cheng, L., Brodsky, R. A., & Biesecker, L. G. (2012). The Phenotype of a Germline Mutation in PIGA: The Gene Somatically Mutated in Paroxysmal Nocturnal Hemoglobinuria. *The American Journal of Human Genetics*, *90*(2), 295–300. https://doi.org/10.1016/j.ajhg.2011.11.031

Jorde, L. B., Hasstedt, S. J., Ritvo, E. R., Mason-Brothers, A., Freeman, B. J., Pingree, C., McMahon, W. M., Petersen, B., Jenson, W. R., & Mo, A. (1991). Complex segregation analysis of autism. *American Journal of Human Genetics*, *49*(5), 932–938.

Kaminsky, E. B., Kaul, V., Paschall, J., Church, D. M., Bunke, B., Kunig, D., Moreno-De-Luca, D., Moreno-De-Luca, A., Mulle, J. G., Warren, S. T., Richard, G., Compton, J. G., Fuller, A. E., Gliem, T. J., Huang, S., Collinson, M. N., Beal, S. J., Ackley, T., Pickering, D. L., … Martin, C. L. (2011). An evidence-based approach to establish

the functional and clinical significance of CNVs in intellectual and developmental disabilities. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *13*(9), 777–784. https://doi.org/10.1097/GIM.0b013e31822c79f9

Kapoor, A., Jiang, Q., Chatterjee, S., Chakraborty, P., Sosa, M. X., Berrios, C., & Chakravarti, A. (2015). Population variation in total genetic risk of Hirschsprung disease from common RET, SEMA3 and NRG1 susceptibility polymorphisms. *Human Molecular Genetics*, *24*(10), 2997–3003. https://doi.org/10.1093/hmg/ddv051

Karakoc, E., Alkan, C., O'Roak, B. J., Dennis, M. Y., Vives, L., Mark, K., Rieder, M. J., Nickerson, D. A., & Eichler, E. E. (2012). Detection of structural variants and indels within exome data. *Nature Methods*, *9*(2), 176–178. https://doi.org/10.1038/nmeth.1810

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kelwick, R., Desanlis, I., Wheeler, G. N., & Edwards, D. R. (2015). The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0676-3

Khayat, M., Tilghman, J. M., Chervinsky, I., Zalman, L., Chakravarti, A., & Shalev, S. A. (2016). A PIGN Mutation Responsible for Multiple Congenital Anomalies–Hypotonia–Seizures Syndrome 1 (MCAHS1) in an Israeli–Arab Family. *American Journal of Medical Genetics. Part A*, *170A*(1), 176–182. https://doi.org/10.1002/ajmg.a.37375

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., & Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics*, *203*(3), 253–310. https://doi.org/10.1002/aja.1002030302

Knight, L. A., Yong, M. H., Tan, M., & Ng, I. S. (1995). Del(3) (p25.3) without phenotypic effect. *Journal of Medical Genetics*, *32*(12), 994–995.

Kowalczyk-Quintas, C., & Schneider, P. (2014). Ectodysplasin A (EDA) – EDA receptor signalling and its pharmacological modulation. *Cytokine & Growth Factor Reviews*, *25*(2), 195–203. https://doi.org/10.1016/j.cytogfr.2014.01.004

Krawitz, P. M., Murakami, Y., Rieß, A., Hietala, M., Krüger, U., Zhu, N., Kinoshita, T., Mundlos, S., Hecht, J., Robinson, P. N., & Horn, D. (2013). PGAP2 Mutations, Affecting the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation Syndrome. *The American Journal of Human Genetics*, *92*(4), 584–589. https://doi.org/10.1016/j.ajhg.2013.03.011

Krawitz, P. M., Schweiger, M. R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., Isau, M., Fischer, A., Dahl, A., Kerick, M., Hecht, J., Köhler, S., Jäger, M., Grünhagen, J., de Condor, B. J., … Robinson, P. N. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nature Genetics*, *42*(10), 827–829. https://doi.org/10.1038/ng.653

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., NHLBI Exome Sequencing Project, Quinlan, A. R., Nickerson, D. A., & Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, *22*(8), 1525–1532. https://doi.org/10.1101/gr.138115.112

Kuhlman, J., & Eisen, J. S. (2007). Genetic screen for mutations affecting development and function of the enteric nervous system. *Developmental Dynamics*, *236*(1), 118–127. https://doi.org/10.1002/dvdy.21033

Kvarnung, M., Nilsson, D., Lindstrand, A., Korenke, G. C., Chiang, S. C. C., Blennow, E., Bergmann, M., Stödberg, T., Mäkitie, O., Anderlid, B.-M., Bryceson, Y. T., Nordenskjöld, M., & Nordgren, A. (2013). A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in *PIGT*. *Journal of Medical Genetics*, *50*(8), 521–528. https://doi.org/10.1136/jmedgenet-2013-101654

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Leoyklang, P., Siriwan, P., & Shotelersuk, V. (2006). A mutation of the p63 gene in non-syndromic cleft lip. *Journal of Medical Genetics*, *43*(6), e28. https://doi.org/10.1136/jmg.2005.036442

Li, C. C., Weeks, D. E., & Chakravarti, A. (1993). Similarity of DNA Fingerprints Due to Chance and Relatedness. *Human Heredity*, *43*(1), 45–52. https://doi.org/10.1159/000154113

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Heng. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Li, Heng, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Wang, Y., Lin, M., Yuan, G., Yang, G., Zheng, Y., & Chen, Y. (2013). Augmented BMPRIA-Mediated BMP Signaling in Cranial Neural Crest Lineage Leads to Cleft Palate Formation and Delayed Tooth Differentiation. *PLoS ONE*, *8*(6). https://doi.org/10.1371/journal.pone.0066107

Lin, R., Tao, R., Gao, X., Li, T., Zhou, X., Guan, K.-L., Xiong, Y., & Lei, Q.-Y. (2013). Acetylation Stabilizes ATP-Citrate Lyase to Promote Lipid Biosynthesis and Tumor Growth. *Molecular Cell*, *51*(4), 506–518. https://doi.org/10.1016/j.molcel.2013.07.002

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B. J., Saucedo-Cardenas, O., Barker, D. F., Killian, J. M., Garcia, C. A., Chakravarti, A., & Patel, P. I. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, *66*(2), 219–232. https://doi.org/10.1016/0092-8674(91)90613-4

Mattioli, F., Schaefer, E., Magee, A., Mark, P., Mancini, G. M., Dieterich, K., Von Allmen, G., Alders, M., Coutton, C., van Slegtenhorst, M., Vieville, G., Engelen, M., Cobben, J. M., Juusola, J., Pujol, A., Mandel, J.-L., & Piton, A. (2017). Mutations in Histone Acetylase Modifier BRPF1 Cause an Autosomal-Dominant Form of Intellectual Disability with Associated Ptosis. *The American Journal of Human Genetics*, *100*(1), 105–116. https://doi.org/10.1016/j.ajhg.2016.11.010

Maydan, G., Noyman, I., Har-Zahav, A., Neriah, Z. B., Pasmanik-Chor, M., Yeheskel, A., Albin-Kaplanski, A., Maya, I., Magal, N., Birk, E., Simon, A. J., Halevy, A., Rechavi, G., Shohat, M., Straussberg, R., & Basel-Vanagaite, L. (2011). Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. *Journal of Medical Genetics*, *48*(6), 383–389. https://doi.org/10.1136/jmg.2010.087114

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Menezes, M., Corbally, M., & Puri, P. (2006). Long-term results of bowel function after treatment for Hirschsprung's disease: A 29-year review. *Pediatric Surgery International*, *22*(12), 987–990. https://doi.org/10.1007/s00383-006-1783-8

More, C. B., Bhavsar, K., Joshi, J., Varma, S. N., & Tailor, M. (2013). Hereditary ectodermal dysplasia: A retrospective study. *Journal of Natural Science, Biology, and Medicine*, *4*(2), 445–450. https://doi.org/10.4103/0976-9668.117012

Mowat, D. R. (2003). Mowat-Wilson syndrome. *Journal of Medical Genetics*, *40*(5), 305–310. https://doi.org/10.1136/jmg.40.5.305

Mulligan, L. M. (2014). RET revisited: Expanding the oncogenic portfolio. *Nature Reviews Cancer*, *14*(3), 173–186. https://doi.org/10.1038/nrc3680

Murphy, D. A., Diaz, B., Bromann, P. A., Tsai, J. H., Kawakami, Y., Maurer, J., Stewart, R. A., Izpisúa-Belmonte, J. C., & Courtneidge, S. A. (2011). A Src-Tks5 Pathway Is Required for Neural Crest Cell Migration during Embryonic Development. *PLoS ONE*, *6*(7), e22499. https://doi.org/10.1371/journal.pone.0022499

Ng, B. G., Hackmann, K., Jones, M. A., Eroshkin, A. M., He, P., Wiliams, R., Bhide, S., Cantagrel, V., Gleeson, J. G., Paller, A. S., Schnur, R. E., Tinschert, S., Zunich, J., Hegde, M. R., & Freeze, H. H. (2012). Mutations in the Glycosylphosphatidylinositol Gene PIGL Cause CHIME Syndrome. *The American Journal of Human Genetics*, *90*(4), 685–688. https://doi.org/10.1016/j.ajhg.2012.02.010

Ohba, C., Okamoto, N., Murakami, Y., Suzuki, Y., Tsurusaki, Y., Nakashima, M., Miyake, N., Tanaka, F., Kinoshita, T., Matsumoto, N., & Saitsu, H. (2014). PIGN mutations cause congenital anomalies, developmental delay, hypotonia, epilepsy, and progressive cerebellar atrophy. *Neurogenetics*, *15*(2), 85–92. https://doi.org/10.1007/s10048-013-0384-7

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., Turner, E. H., Stanaway, I. B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J. M., Borenstein, E., Rieder, M. J., … Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, *485*(7397), 246–250. https://doi.org/10.1038/nature10989

Passarge, E. (1967). The Genetics of Hirschsprung's Disease. *New England Journal of Medicine*, *276*(3), 138–143. https://doi.org/10.1056/NEJM196701192760303

Pohjola, P., Leeuw, N. de, Penttinen, M., & Kääriäinen, H. (2010). Terminal 3p deletions in two families—Correlation between molecular karyotype and phenotype. *American Journal of Medical Genetics Part A*, *152A*(2), 441–446. https://doi.org/10.1002/ajmg.a.33215

Presson, A. P., Partyka, G., Jensen, K. M., Devine, O. J., Rasmussen, S. A., McCabe, L. L., & McCabe, E. R. B. (2013). Current Estimate of Down Syndrome Population Prevalence in the United States. *The Journal of Pediatrics*, *163*(4), 1163–1168. https://doi.org/10.1016/j.jpeds.2013.06.013

Puffenberger, E. G., Hosoda, K., Washington, S. S., Nakao, K., deWit, D., Yanagisawa, M., & Chakravarti, A. (1994). A missense mutation of the endothelin-B receptor gene in multigenic hirschsprung's disease. *Cell*, *79*(7), 1257–1266. https://doi.org/10.1016/0092-8674(94)90016-7

Pummila, M., Fliniaux, I., Jaatinen, R., James, M. J., Laurikkala, J., Schneider, P., Thesleff, I., & Mikkola, M. L. (2007). Ectodysplasin has a dual role in ectodermal organogenesis: Inhibition of Bmp activity and induction of Shh expression. *Development*, *134*(1), 117–125. https://doi.org/10.1242/dev.02708

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Ray, A. K., Marazita, M. L., Pathak, R., Beever, C. L., Cooper, M. E., Goldstein, T., Shaw, D. F., & Field, L. L. (2004). TP63 mutation and clefting modifier genes in an EEC syndrome family. *Clinical Genetics*, *66*(3), 217–222. https://doi.org/10.1111/j.1399-0004.2004.00287.x

Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F., & Ronald, A. (2013). Examining and interpreting the female protective effect against autistic behavior. *Proceedings of the National Academy of Sciences*, *110*(13), 5258–5262. https://doi.org/10.1073/pnas.1211070110

Sabari, B. R., Tang, Z., Huang, H., Yong-Gonzalez, V., Molina, H., Kong, H. E., Dai, L., Shimada, M., Cross, J. R., Zhao, Y., Roeder, R. G., & Allis, C. D. (2015). Intracellular Crotonyl-CoA Stimulates Transcription Through p300-Catalyzed Histone Crotonylation. *Molecular Cell*, *58*(2), 203–215. https://doi.org/10.1016/j.molcel.2015.02.029

Sadier, A., Viriot, L., Pantalacci, S., & Laudet, V. (2014). The ectodysplasin pathway: From diseases to adaptations. *Trends in Genetics: TIG*, *30*(1), 24–31. https://doi.org/10.1016/j.tig.2013.08.006

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., Murtha, M. T., Bal, V. H., Bishop, S. L., Dong, S., Goldberg, A. P., Jinlu, C., Keaney, J. F., Klei, L., Mandell, J. D., Moreno-De-Luca, D., Poultney, C. S., Robinson, E. B., Smith, L., … State, M. W. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, *87*(6), 1215–1233. https://doi.org/10.1016/j.neuron.2015.09.016

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., … State, M. W. (2012). De novo mutations revealed by whole exome sequencing are strongly associated with autism. *Nature*, *485*(7397), 237–241. https://doi.org/10.1038/nature10945

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Scapoli, L., Martinelli, M., Arlotti, M., Palmieri, A., Masiero, E., Pezzetti, F., & Carinci, F. (2008). Genes causing clefting syndromes as candidates for non-syndromic cleft lip with or without cleft palate: A family-based association study. *European Journal of Oral Sciences*, *116*(6), 507–511. https://doi.org/10.1111/j.1600-0722.2008.00574.x

Seuntjens, E., Nityanandam, A., Miquelajauregui, A., Debruyn, J., Stryjewska, A., Goebbels, S., Nave, K.-A., Huylebroeck, D., & Tarabykin, V. (2009). Sip1 regulates sequential fate decisions by feedback signaling from postmitotic neurons to progenitors. *Nature Neuroscience*, *12*(11), 1373–1380. https://doi.org/10.1038/nn.2409

Shi, H., Enriquez, A., Rapadas, M., Martin, E. M. M. A., Wang, R., Moreau, J., Lim, C. K., Szot, J. O., Ip, E., Hughes, J. N., Sugimoto, K., Humphreys, D. T., McInerney-Leo, A. M., Leo, P. J., Maghzal, G. J., Halliday, J., Smith, J., Colley, A., Mark, P. R., … Dunwoodie, S. L. (2017). NAD Deficiency, Congenital Malformations, and Niacin Supplementation. *New England Journal of Medicine*, *377*(6), 544–552. https://doi.org/10.1056/NEJMoa1616361

Shuib, S., McMullan, D., Rattenberry, E., Barber, R. M., Rahman, F., Zatyka, M., Chapman, C., Macdonald, F., Latif, F., Davison, V., & Maher, E. R. (2009). Microarray based analysis of 3p25-p26 deletions (3p- syndrome). *American Journal of Medical Genetics Part A*, *149A*(10), 2099–2105. https://doi.org/10.1002/ajmg.a.32824

Stanchina, L., Baral, V., Robert, F., Pingault, V., Lemort, N., Pachnis, V., Goossens, M., & Bondurand, N. (2006). Interactions between Sox10, Edn3 and Ednrb during enteric nervous system and melanocyte development. *Developmental Biology*, *295*(1), 232–249. https://doi.org/10.1016/j.ydbio.2006.03.031

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D., & Cooper, D. N. (2014). The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, *133*(1), 1–9. https://doi.org/10.1007/s00439-013-1358-4

Suzuki, K., Hu, D., Bustos, T., Zlotogora, J., Richieri-Costa, A., Helms, J. A., & Spritz, R. A. (2000). Mutations of PVRL1 , encoding a cell-cell adhesion molecule/herpesvirus receptor, in cleft lip/palate-ectodermal dysplasia. *Nature Genetics*, *25*(4), 427–430. https://doi.org/10.1038/78119

Szot, J. O., Campagnolo, C., Cao, Y., Iyer, K. R., Cuny, H., Drysdale, T., Flores-Daboub, J. A., Bi, W., Westerfield, L., Liu, P., Leung, T. N., Choy, K. W., Chapman, G., Xiao, R., Siu, V. M., & Dunwoodie, S. L. (2020). Bi-allelic Mutations in NADSYN1 Cause Multiple Organ Defects and Expand the Genotypic Spectrum of Congenital NAD Deficiency Disorders. *The American Journal of Human Genetics*, *106*(1), 129–136. https://doi.org/10.1016/j.ajhg.2019.12.006

Takagishi, J., Rauen, K. A., Drumheller, T., Kousseff, B., & Sutcliffe, M. (2006). Chromosome 3p25 deletion in mother and daughter with minimal phenotypic effect. *American Journal of Medical Genetics Part A*, *140A*(14), 1587–1593. https://doi.org/10.1002/ajmg.a.31325

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Tossell, K., Andreae, L. C., Cudmore, C., Lang, E., Muthukrishnan, U., Lumsden, A., Gilthorpe, J. D., & Irving, C. (2011). Lrrn1 is required for formation of the midbrain–hindbrain boundary and organiser through regulation of affinity differences between midbrain and hindbrain cells in chick. *Developmental Biology*, *352*(2–10), 341–352. https://doi.org/10.1016/j.ydbio.2011.02.002

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression

analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Turner, T. N., Sharma, K., Oh, E. C., Liu, Y. P., Collins, R. L., Sosa, M. X., Auer, D. R., Brand, H., Sanders, S. J., Moreno-De-Luca, D., Pihur, V., Plona, T., Pike, K., Soppet, D. R., Smith, M. W., Cheung, S. W., Martin, C. L., State, M. W., Talkowski, M. E., … Chakravarti, A. (2015). Loss of δ-catenin function in severe autism. *Nature*, *520*(7545), 51–56. https://doi.org/10.1038/nature14186

Van de Putte, T., Maruhashi, M., Francis, A., Nelles, L., Kondoh, H., Huylebroeck, D., & Higashi, Y. (2003). Mice Lacking Zfhx1b, the Gene That Codes for Smad-Interacting Protein-1, Reveal a Role for Multiple Neural Crest Cell Defects in the Etiology of Hirschsprung Disease–Mental Retardation Syndrome. *American Journal of Human Genetics*, *72*(2), 465–470.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. https://doi.org/10.1093/nar/gkq603

Werling, D. M., & Geschwind, D. H. (2015). Recurrence rates provide evidence for sex-differential, familial genetic liability for autism spectrum disorders in multiplex families and twins. *Molecular Autism*, *6*(1), 27. https://doi.org/10.1186/s13229-015-0004-5

Westerfield, M. (1991). *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish Danio (Brachydanio) rerio* (4th ed.). University of Oregon Press. https://norecopa.no/textbase/a-guide-for-the-laboratory-use-of-zebrafish-danio-brachydanio-rerio

Wright, C. F., FitzPatrick, D. R., & Firth, H. V. (2018). Paediatric genomics: Diagnosing rare disease in children. *Nature Reviews Genetics*, *19*(5), 253–268. https://doi.org/10.1038/nrg.2017.116

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N., & Tyler-Smith, C. (2012). Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. *American Journal of Human Genetics*, *91*(6), 1022–1032. https://doi.org/10.1016/j.ajhg.2012.10.015

Yang, A., Schweitzer, R., Sun, D., Kaghad, M., Walker, N., Bronson, R. T., Tabin, C., Sharpe, A., Caput, D., Crum, C., & McKeon, F. (1999). P63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. *Nature*, *398*(6729), 714–718. https://doi.org/10.1038/19539

Zhang, Y., Tomann, P., Andl, T., Gallant, N. M., Huelsken, J., Jerchow, B., Birchmeier, W., Paus, R., Piccolo, S., Mikkola, M. L., Morrisey, E. E., Overbeek, P. A., Scheidereit, C., Millar, S. E., & Schmidt-Ullrich, R. (2009). Reciprocal requirements for Eda/Edar/NF-κB and Wnt/β-catenin signaling pathways in hair follicle induction. *Developmental Cell*, *17*(1), 49–61. https://doi.org/10.1016/j.devcel.2009.05.011

Zhang, Z., & Henzel, W. J. (2004). Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Science*, *13*(10), 2819–2824. https://doi.org/10.1110/ps.04682504

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Permissions

Chapter 2:

Section 2.3 of Chapter 2 is reprinted from the journal article:

> Khayat, M., Tilghman, J. M., Chervinsky, I., Zalman, L., Chakravarti, A., & Shalev, S. A. (2016). A PIGN Mutation Responsible for Multiple Congenital Anomalies–Hypotonia–Seizures Syndrome 1 (MCAHS1) in an Israeli–Arab Family. *American Journal of Medical Genetics. Part A*, *170A*(1), 176–182. https://doi.org/10.1002/ajmg.a.37375

Reprint permissions for this article follow:

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

May 31, 2020

This Agreement between Johns Hopkins University -- Joseph Tilghman ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4838560608724
License date	May 29, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	American Journal of Medical Genetics Part A
Licensed Content Title	A PIGN mutation responsible for multiple congenital anomalies–hypotonia–seizures syndrome 1 (MCAHS1) in an Israeli–Arab family
Licensed Content Author	Morad Khayat, Joseph Mark Tilghman, Ilana Chervinsky, et al
Licensed Content Date	Sep 14, 2015

Licensed Content Volume	170
Licensed Content Issue	1
Licensed Content Pages	7
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title	Sequence analysis of familial neurodevelopmental disorders
Institution name	Johns Hopkins University
Expected presentation date	Jul 2020
Order reference number	1
Requestor Location	Johns Hopkins University 733 N Broadway MRB 515 BALTIMORE, MD 21205 United States Attn: Johns Hopkins University
Publisher Tax ID	EU826007151
Total	0.00 USD

Chapter 3:

The following statement applies to the entirety of Chapter 3, for which no formal reprinting permission is required (see https://www.nejm.org/about-nejm/permissions):

Reproduced with permission from (Tilghman, J. M., Ling, A. Y., Turner, T. N., Sosa, M. X., Krumm, N., Chatterjee, S., Kapoor, A., Coe, B. P., Nguyen, K.-D. H., Gupta, N., Gabriel, S., Eichler, E. E., Berrios, C., & Chakravarti, A. (2019). Molecular Genetic Anatomy and Risk Profile of Hirschsprung's Disease. New England Journal of Medicine, 380(15), 1421–1432. https://doi.org/10.1056/NEJMoa1706594), Copyright Massachusetts Medical Society.

CURRICULUM VITAE

Joseph Tilghman

PhD candidate in the Aravinda Chakravarti Lab
New York University Langone Medical Center
josephtilghman@gmail.com
(443)944-1036

EDUCATION

Ph.D., Human Genetics
Advisor: Aravinda Chakravarti, Ph.D.
Johns Hopkins School of Medicine *December 2020*

B.S., Biology
Washington and Lee University *2013*

JOURNAL ARTICLES

Mederer, T., Schmitteckert, S., Volz, J., Martínez, C., Röth, R., Thumberger, T., Eckstein, V., Scheuerer, J., Thöni, C., Lasitschka, F., Carstensen, L., Günther, P., Holland-Cunz, S., Hofstra, R., Brosens, E., Rosenfeld, J. A., Schaaf, C. P., Schriemer, D., Ceccherini, I., Rusmini, M., **Tilghman, J.,** … Niesler, B. (2020). A complementary study approach unravels novel players in the pathoetiology of Hirschsprung disease. *PLOS Genetics*, *16*(11), e1009106. https://doi.org/10.1371/journal.pgen.1009106

Tilghman, J. M., Ling, A. Y., Turner, T. N., Sosa, M. X., Krumm, N., Chatterjee, S., Kapoor, A., Coe, B. P., Nguyen, K.-D. H., Gupta, N., Gabriel, S., Eichler, E. E., Berrios, C., & Chakravarti, A. (2019). Molecular Genetic Anatomy and Risk Profile of Hirschsprung's Disease. *New England Journal of Medicine*, *380*(15), 1421–1432. https://doi.org/10.1056/NEJMoa1706594

Marsh, D. M., Townes, F. W., Cotter, K. M., Farroni, K., McCreary, K. L., Petry, R. L., & **Tilghman, J. M.** (2019). Thermal Preference and Species Range in Mountaintop Salamanders and Their Widespread Competitors. *Journal of Herpetology*, *53*(2), 96–103. https://doi.org/10.1670/18-110

Khayat, M., **Tilghman, J. M.**, Chervinsky, I., Zalman, L., Chakravarti, A., & Shalev, S. A. (2016). A PIGN Mutation Responsible for Multiple Congenital Anomalies–Hypotonia–Seizures Syndrome 1 (MCAHS1) in an Israeli–Arab Family. *American Journal of Medical Genetics. Part A*, *170A*(1), 176–182. https://doi.org/10.1002/ajmg.a.37375

Watson, F. L., Schmidt, H., Turman, Z. K., Hole, N., Garcia, H., Gregg, J., **Tilghman, J.**, & Fradinger, E. A. (2014). Organophosphate pesticides induce morphological abnormalities and decrease locomotor activity and heart rate in Danio rerio and Xenopus laevis. *Environmental Toxicology and Chemistry*, *33*(6), 1337–1345. https://doi.org/10.1002/etc.2559

Tilghman, J. M., Ramee, S. W., & Marsh, D. M. (2012). Meta-analysis of the effects of canopy removal on terrestrial salamander populations in North America. *Biological Conservation*, *152*, 1–9. https://doi.org/10.1016/j.biocon.2012.03.030

PRESENTATIONS

Tilghman, J. M., Ling, A. Y., Turner, T. N., Sosa, M. X., Krumm, N., Chatterjee, S., Kapoor, A., Coe, B. P., Nguyen, K. D. H., Gupta, N., Gabriel, S., Eichler, E. E., Berrios, C., & Chakravarti, A. (2017, October). *The molecular genetic anatomy and risk profile of Hirschsprung disease.* Invited lecture at Washington and Lee University Biology Department, Lexington, VA.

Tilghman, J. M., Ling, A. Y., Turner, T. N., Sosa, M. X., Krumm, N., Chatterjee, S., Kapoor, A., Coe, B. P., Nguyen, K. D. H., Gupta, N., Gabriel, S., Eichler, E. E., Berrios, C., & Chakravarti, A. (2017, May). *A molecular genetic anatomy of Hirschsprung disease and its risk profile.* Invited lecture at meeting of the International Hirschsprung Disease Consortium and poster presented at the 50th Annual Meeting of the European Society of Human Genetics, Copenhagen, Denmark.

Tilghman J. M., Stevens C. R., Daly M. J., Chakravarti, A. (2015, October). *Autism gene discovery in rare female-enriched multiplex families (FEMFs).* Poster presented at the 65th Annual Meeting of The American Society of Human Genetics, Baltimore, MD.

RESEARCH EXPERIENCES

Sequence Analysis of Familial Neurodevelopmental Disorders
Aravinda Chakravarti Lab *2014 — Present*

- Analyzed exome sequencing on probands from consanguineous Israeli Arab families in order to identify etiology of neurodevelopmental syndromes
- Assessed pathogenicity of rare coding variation, genomic copy number variation, and common risk genotypes for Hirschsprung disease and combined these variant classes to estimate individual genetic risk on the basis of observed genotypes using a cohort of affecteds and ancestry matched controls
    - Compiled and assessed all literature reported pathogenic variation for Hirschsprung disease and Hirschsprung associated syndromes
    - Wrote rare variant annotation programs based on literature review for well characterized genes in order to compare general annotation based

methods of variant pathogenicity determination to gene-specific methods
    - Completed rigorous quality control and analysis of exome sequencing and other genotypic datasets
    - Jointly analyzed heterogeneous variant datasets and clinical data in order to describe the molecular genetic risk profile of Hirshsprung disease, allowing for quantitative assessment of individual disease risk
- Performed exome sequencing analysis of a cohort of autism families characterized by high genetic risk
    - Reviewed evidence supporting reported syndromic autism genes in order to make a list of syndromes having high confidence autism association
    - Completed calling and analysis of multiple local and public highly parallel sequencing datasets including thousands individuals, including ancestry determination, sex checks, relatedness checks, etc. in order to compare the genetic risk profiles of families having high and low genetic risk
    - Assessed enrichment of rare coding variation in syndromic and idiopathic autism genes
    - Identified candidate causative genes for families on the basis of transmission and enrichment of likely pathogenic variation compared to controls
- Worked with genetic counselors to assess pathogenicity of clinically identified variants based genetic findings from our lab and information within public databases
- Analyzed and advised others on analysis of exome sequencing, genome-wide genotyping, genomic assays, etc. in support of various lab projects

Undergraduate Research in Neurodevelopmental Teratology
Fiona Watson Lab *2013*

- Oversaw and carried out behavioral and morphological assays to determine the effect of the organophosphate pesticide chlorpyrifos on *Xenopus laevis* embryo development
    - Directed student performed behavioral assays carried out by a developmental biology class
    - Compiled and analyzed assay results

Undergraduate Research Fellowship in Poplar Genetics
Stephen DiFazio Lab *Summer 2012*

- Performed multiplexed PCR assays in order to validate a set of genomic insertions and deletions in poplar trees, previously identified by whole genome sequencing
    - Validated primer sets and performed PCR-based assays
    - Performed sanger sequencing as a means of confirming assay results
    - Analyzed the results of PCR assays with respect to called genotypes from whole genome sequencing and expected genotypic ratios

Undergraduate Research in Ecology
David Marsh Lab *2010 — 2013*

- Performed literature review and meta-analysis of the effects of forestry on terrestrial salamander abundance
- Supervised and carried out field work over a two year interval as part of a study of the differential behavior of salamander species with overlapping distributions as it relates to climate

TEACHING AND MENTORING EXPERIENCE

Analysis of Rare Variants Involved in Syndromic Autism
Research Mentor *Summer 2017*

- Designed a summer research project for and served as a mentor to a recent college graduate as part of a summer research fellowship for underrepresented minority trainees, funded by the Simons Foundation Autism Research Initiative
- Oversaw her review and analysis of rare coding variation patterns reported as causal for autism associated syndromes
- Guided her in preparing oral presentations on the project for both lab meeting and a poster symposium

Principles of Genetics
Teaching Assistant *Fall 2016 and 2017*

- This course covers a range of topics in genetics for graduate students from several departments in the Johns Hopkins School of Medicine
- Defined learning objectives, co-developed in-class lectures, created online lectures and other supplementary materials, held review sessions and office hours, and developed and graded assessments

Basic Mechanisms of Disease
Teaching Assistant *Fall 2015*

- This course serves as the primary course in human pathobiology for graduate students from several departments and for clinical fellows in genetics
- Gave lectures reviewing the anatomy and physiology of the renal and muscular systems (in preparation for pathology lectures)
- Wrote and graded assessments relating to the pathobiology of the renal and muscular systems

Cell Biology Laboratory
Teaching Assistant *Winter and Fall 2012*

- This is research-based laboratory component of an undergraduate cell biology course
- Helped students with laboratory procedures (e.g. DNA purification, PCR, gene cloning using plasmid vectors, imaging gels, gel extractions, sectioning retinal tissue using a cryostat, immunostaining tissue sections and visualizing them using epifluorescent microscopy, pouring plates)
- Prepared solutions and calibrated equipment

General Biology Laboratory
Teaching Assistant *Fall 2010 and 2011, Winter and Fall 2012*

- This is an inquiry-based standalone laboratory course for undergraduate biology majors
- Assisted students with field work, laboratory procedures, scientific writing and study design, and use of descriptive statistics

SERVICE

American Society of Human Genetics, abstract reviewer *2019*
Human Genetics Program, recruitment committee member *2015 — 2018*
Washington and Lee physiology faculty search, student committee member *2012*

VOLUNTEERING WORK

Project Bridge (science outreach organization), educator *2014 — 2019*
Baltimore Brain Fest, planning committee member *2016 — 2018*

PROFESSIONAL MEMBERSHIPS

American Society of Human Genetics *2014 — Present*
European Society of Human Genetics *2017 — Present*

ACHIEVEMENTS AND HONORS

55th Short Course on Medical & Experimental Mammalian Genetics *2014*
NSF REU Fellowship at West Virginia University *2012*
Howard Hughes Medical Institute Undergraduate Research Fellowship *2010 — 2012*
Delmarva Honors Scholarship (full tuition academic scholarship) *2009 — 2013*
National Merit Scholarship (institutional) *2009 — 2013*
Robert C. Byrd Honors Scholarship (state administered national program) *2009 — 2012*

SKILLS

Genetics/Genomics Analysis Tools: GATK | Plink | Samtools | ANNOVAR | VCFtools | SNPRelate | etc.
Analyses: Extensive experience in development and execution of full next-generation sequencing pipelines (fastq to gene finding) including all aspect of data cleaning | Analysis of array-based genotyping | Determination of individual specific risk on the basis of heterogeneous genetic and other risk factors | Pedigree-based Mendelian gene finding | Use of SLURM and SGE for implementation of highly parallel analysis pipelines for analysis of very large datasets | Sanger Sequencing | Exome and other targeted sequencing assays | Genetic determination of ancestry and relatedness | Determination of variant pathogenicity in a general and gene aware manner | Transmission based gene finding | Association-based gene finding | Developed simulations for association testing and power analyses | Analysis of RNA-seq and other genomics assays | Meta-analysis
Databases/Resources: Gnomad | Clinvar | HGMD | UniProt | OMIM | UCSC Browser | etc.
Programming Languages: Python (primary – 8 years) | R (as needed for statistical analyses and visualization – 6 years) | Bash (work daily in Unix environment on HPC cluster – 7 years)
Research Areas: Human statistical and population genetics | Human Mendelian genetics | Molecular genetics